IYKYK: Using language models to decode extremist cryptolects

Christine de Kock[*]
University of Melbourne

Arij Riabi
INRIA Paris[†]

Zeerak Talat
University of Edinburgh

Michael Sejr Schlichtkrull
Queen Mary University of London

Pranava Madhyasta
City St George's, University of London

Eduard Hovy
University of Melbourne

Abstract

Extremist groups develop complex in-group language, also referred to as *cryptolects*, to exclude or mislead outsiders. We investigate the ability of current language technologies to detect and interpret the cryptolects of two online extremist platforms. Evaluating eight models across six tasks, our results indicate that general purpose large language models cannot consistently detect or decode extremist language. However, performance can be significantly improved by domain adaptation and specialised prompting techniques. These results provide important insights to inform the development and deployment of automated moderation technologies. We further develop and release novel labelled and unlabelled datasets, including 19.4M posts from extremist platforms and lexicons validated by human experts.

1 Introduction

The proliferation of online extremist ideologies, often veiled in complex cryptolects, poses a severe societal threat and a unique challenge for Natural Language Processing (NLP). Although online spaces have been used to facilitate the activities of extremist groups since shortly after the creation of the world wide web (Hoffman, 1997; Conway, 2006), concerns over the recent rise of online extremist activity have been expressed by researchers (Baele et al., 2021), the media (Rich and Bujalka, 2023), NGOs (Latimore and Coyne, 2023), and governments (PMO, 2025) alike.

While modelling hate speech and toxic content are well-studied tasks within NLP, the related problem of modelling extremist language has received comparatively little attention. The latter has distinctive characteristics that make it challenging: the vernaculars of different ideologies are very specialised, often subtle, and highly dynamic (Weimann and

---
[*]**Correspondence:** christine.dekock@unimelb.edu.au
[†]Now at Devoteam

Figure 1: Three posts from the IYKYK dataset (in gray) with translations by `Llama-3.3` (in green). The posts originate from Incels (#1 & #2) and Stormfront (#3).

Am, 2020). Bogetić (2023) states that "the neologisms being produced by Incels have come to form a true cryptolect, developing at a rate that almost escapes linguistic description". In addition to neologisms, extremist groups also create "dogwhistles" by imbuing existing words with a coded meaning unknown to the general population (Mendelsohn et al., 2023). Furthermore, all extremist language is not necessarily hateful: specialised language may be used to signal in-group status while discussing benign topics (McCulloch, 2019), meaning that sentence context is not always helpful.

The result is that, although these groups often interact on public discussion forums, their language

can be incomprehensible to outsiders. Furthermore, it is not sufficient to be able to detect extremist language: being able to decode and interpret it is essential. Fig. 1 illustrates this problem, showing three posts with translations by Llama 3.3 (Grattafiori et al., 2024). In the first post, we note the phrase *going ER*, which refers to committing a mass murder targeting women[1], and the term *roping*, which refers to suicide. Both terms are inaccurately decoded by the model, illustrating the importance of understanding the limitations of these models. While methods and datasets have been proposed to identify users at risk of radicalization (De Kock and Hovy, 2024) and to detect extremist content (Riabi et al., 2025), the question remains of how to decode and interpret the cryptolects developed by these groups.

Complicating matters, toxic language is typically removed during pretraining of modern NLP models (Anil et al., 2023; Longpre et al., 2024) to prevent the generation of toxic content. Removing such data can even improve performance on non-toxic domains (Soldaini et al., 2024). This prevents models from generating toxic content, but could limit their ability to model extremist cryptolects.

The central hypothesis of this work is that generalised language technologies require specialised techniques when applied to extremist cryptolects. We evaluate this hypothesis in two ways. Firstly, we investigate the ability of instruction-tuned models to recognise and interpret extremist language in a zero-shot setting. Secondly, we evaluate the impact of domain adaptation for supervised classification with fine-tuned encoder models. To enable these experiments, we create a dataset[2] of 19.4M posts from two online extremist forums, Stormfront and Incels, that are dedicated to white supremacist and misogynistic ideologies, respectively. We further create test sets for three tasks using this data, employing human experts for annotation.

These test sets are used in our zero-shot evaluation to compare seven instruction-tuned large language models (LLMs). For the task of identifying in-group language, the top models achieve an F1 of 64.4% and 80.0% for Stormfront and Incels, respectively. For both platforms, adding context (in the form of post examples) increases recall while

decreasing precision, while adding more extensive instructions yields an F1 increase of 10%. For the task of generating definitions, adding post examples improves performance for both platforms, with 90.3% of definitions being correct for Incels when including 10 posts, while only 44.7% are correctly generated when no context is provided. All models perform better at identifying and defining in-group language for Incels compared to Stormfront, which is impacted by the former's tendency to reuse common morphemes when creating new words.

For detecting radical content, we find that the domain adaptation of encoder LLMs consistently improves performance across two tasks, achieving state-of-the-art results in English. In particular, fine-tuning xlmt on our dataset significantly boosts macro-F1 scores (+3.6% and +5.2%), especially when combined with auxiliary task such as ideology prediction. Interestingly, the results indicate that performance is improved even when the domain adaptation uses data from an unrelated extremist ideology. Finally, we find that domain adaptation on our data also improves the performance of classifiers for hate speech detection.

2 Background

2.1 Hateful in-group language

Our work is grounded in two recent works in NLP that investigate the use of specialised language to convey a specific (often hateful) message to an in-group. We detail their approaches here.

Mendelsohn et al. (2023) investigate the topic of dogwhistles using a newly-created glossary of 300 antisemetic dogwhistles. They evaluate whether an LLM can correctly provide dogwhistles' covert meanings, finding that providing additional cues in the prompt (such as the definition of the concept of a dogwhistle) improves performance. Secondly, they investigate the ability to surface dogwhistles, prompting the model to generate examples based on the definition and comparing the output with their glossary. Finally, they perform a small-scale manual analysis of in-context dogwhistle recognition over 5 samples, prompting the model to identify dogwhistles given a real-world example usage. They note that a larger-scale, more rigorous analysis is warranted for the latter task. In this work, we build on their analysis, experimenting with seven models, three validation tasks, and four levels of context over two communities.

Secondly, de Kock (2025) introduces an ap-

---

[1] ER refers to Elliot Rodger, a mass murderer and icon within the incel community.

[2] All models and datasets produced in this work will be made available to verified researchers upon request at this link.

proach for inducing lexicons of extremist in-group language based on social and temporal characteristics of online conversations. Manual validation by experts is used to create a test set. Given a word and usage examples, experts are tasked with assigning a binary label representing whether the word constitutes in-group language of the group in question. In this work, we evaluate the ability of LLMs to replicate this task, which is taxing on human moderators. We use their lexicon induction and validation approaches as a basis for finding candidate words. While their approach relied primarily on social and statistical cues, we focus on language models in our evaluation.

To our knowledge, our study is the first to perform an in-depth exploration of the ability of LLMs to identify and decode extremist in-group language.

2.2 Detecting radical content

Online extremism is frequently fuelled by exposure to radical content (Valentini, 2021; Shaw, 2023). The availability of datasets for detecting radical content remains limited, both in terms of the range of ideologies and the representation of linguistic diversity (Gaikwad et al., 2023; Berjawi et al., 2023).

One major challenge stems from the difficulty of clearly defining what constitutes extremism or radicalization, which is a long-debated topic within counterterrorism research (Della Porta and LaFree, 2012). These ambiguities make annotation particularly challenging and can introduce biases in how datasets are constructed (Gaikwad et al., 2021). In this work, we use the term *extremism* to refer to ideologies that promote violence against an out-group. Following Riabi et al. (2025), *radical content* is understood to refer to texts and other media expressing a radical perspective in opposition to a political, social, or religious system.

The recently introduced Counter dataset (Riabi et al., 2025) provides annotations of radical content in English, French, and Arabic, representing primarily Jihadist and alt-right ideologies. In Riabi et al. (2025), pretrained encoder models are fine-tuned to perform classification on a range of tasks, such as inferring the radicalization level and calls for action. However, the concepts underlying radicalization remain sensitive to the language used, the specific ideological context, and the particular forums in which the content is produced. In this work, we use two tasks in the Counter dataset to investigate the effect of language variation on pretrained encoder models.

2.3 Cryptolects, slang, and jargon

Cryptolects are closely related to other forms of informal language, including slang and jargon. Slang refers to informal neologisms more broadly, whereas jargon refers to technical terminology in professional groups (Mattiello et al., 2005). They may overlap with cryptolects, but there are also important distinctions (Hancock, 1986). First, cryptolects have a stronger focus on obfuscating communication from outsiders, which makes it more challenging to detect and interpret. Second, slang often becomes adopted widely into society due its ability to present a word to convey a concept (e.g., "woke" as to convey being aware of systemic oppression) which may be missing from existing vocabulary. Cryptolects, on the other hand, do not generally share this feature, and are thus less likely to see wider adoption.

Prior work has investigated these related phenomena; for instance, the task definitions of (Sun et al., 2024) and Lucy et al. (2023) relate to ours in sharing the goal of identifying community-specific language. Our work extends beyond this by also seeking for models to define it, to investigate not just the pattern recognition abilities, but the knowledge and expectation for models to correctly identify and delineate the words outside of typical contexts. Furthermore, we also propose domain adaptation as a solution to the shortcomings of pretrained models on such language.

3 The IYKYK dataset

	Incels	Stormfront
Start date	2017-11-08	2001-09-11
# posters in 2024	3827	1784
# months of data	85	279
# posts	10.3 MM	9.3 MM
# posters	16 654	97 430
# posts / posters	25	3
# posts / thread	14	5
# char / post	44	185

Table 1: Characteristics of the dataset. Where relevant, medians are used to calculate averages.

To enable our zero-shot and domain adaptation experiments, we construct a new dataset of posts from Incels and Stormfront. The dataset consists of all publicly available posts on both platforms up to November 2024, totalling 19.4 million posts. For each post, we include thread ID, platform subcategory, user ID, and posting date as meta data. Our dataset only includes users who have posted

on the platforms and therefore does not reflect the total number of users or visitors of each platform[3].

We note that there are differences between the two communities in our dataset (see Table 1). Specifically, users on Incels tend to be more active, with each user posting an average of 25 posts, compared to 3 on Stormfront. Posts on Stormfront, however, are on average 4.2 times longer than on Incels, while Incels features longer threads (2.8x). Using the `langid` package, we find that the posts are predominantly in English (95% for Stormfront and 87% for Incels) with minor French, Spanish, and German components in both; however, we expect that there is some noise in this result due to the non-standard nature of the language.

4 Experimental setting

We examine two different settings to evaluate the applicability of current dominant language technologies to extremism-related tasks. In the first part, we investigate the utility of zero-shot evaluation of seven instruction-tuned models for *classifying* in-group language, *retrieving* instances of in-group language, and *decoding* (i.e., defining) in-group terminology. In the second part, we perform supervised classification, using `xlmt` to evaluate the *domain adaptation* capabilities of large pre-trained language models for detecting radical content.

4.1 Zero-shot evaluation

Fig. 2 illustrates the tasks we use to evaluate the instruction-tuned models. Seven models are used: `Gemma 2` (27B and 9B parameters, Gemma Team et al., 2024), `Llama 3.1` and `3.3` (8B and 72B, Grattafiori et al., 2024), `Mixtral 8x` (7B, Jiang et al., 2024), and `Qwen 2.5` (7B and 72B, Bai et al., 2023). For each task, we experiment with four context framing strategies in the prompt, as detailed below. In Sec. 5, we describe the pipeline for generating test cases and the human experts employed to assign gold standard labels. Full prompts and implementation details are provided in App. C.

Task 1: Classification Given a target word, each model is tasked with evaluating whether it constitutes valid in-group language for a particular community. Here, we closely emulate the human validation setup as described in de Kock (2025).
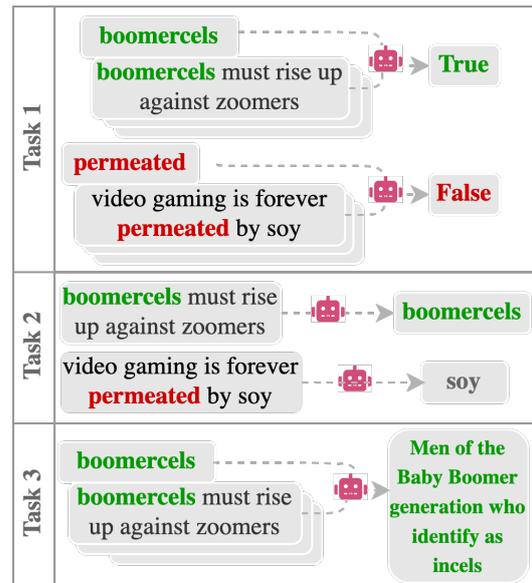

Figure 2: Tasks used for the zero-shot LLM evaluation.

For each platform, 3050 test cases, including positive and negative examples, are evaluated. We report the F1 scores.

Task 2: Retrieval Task 2 requires the model to identify all in-group language within a given post, contextualizing a specific target word (e.g., from the relevant platform). This is more demanding than Task 1 due to the higher ratio of standard to specialized words in typical posts. However, evaluation strictly focuses on the target word: we evaluate if known positive targets are included in the model's output and confirmed negative targets are excluded. For instance (see Fig. 2, Task 2), if *permeated* (a negative sample) is the target, and the model flags *soy* as in-group, our evaluation only checks that *permeated* is not flagged. While similar to the retrieval task of Mendelsohn et al. (2023), our approach enables larger-scale investigation and incorporates confirmed negatives, a need they identified. We again use F1 score as metric.

Task 3: Decoding Given a target word, the model is tasked with providing its definition. For this task, we only evaluate confirmed positive instances of in-group language. Definitions are generated by the instruction-tuned `Llama-3.3` (70B) model. For each platform and context framing, 300 test cases are evaluated, totalling 2400 definitions. Human experts assess the validity of the generated definitions using the labels *correct*, *partially correct*, and *incorrect*. The full instructions are provided in App. B. Test cases are drawn from the

---

positive samples in Tasks 1 and 2 as well as prior lexica by De Kock and Hovy (2024) and de Kock (2025) , aiming to approximately match frequency distributions of the included words between the platforms. We report as metric the validity distribution for each setting.

Context framing We explore four levels of contextual input in the prompts. The `definition` framing parallels the setting of Mendelsohn et al. (2023), providing only the definition for in-group language and the target word. This probes whether the model training data included sufficient information for the model to perform the task without further context. The `instructions` framing provides the full instructions as used for the human annotation (detailed in Sec. 5) which also includes the nature of the research (focusing on extremism), the platform name and ideology, and examples of positive and negative types. The `10-examples` framing provides the full instructions and ten posts containing the target word on the relevant platform. Since there is a disparity in the post lengths between the two platforms, we select only posts with 10-300 characters. Finally, the `1-example` framing provides full instructions and one example post only. For Task 2, providing examples is infeasible, since a target word is to be retrieved from a given post. As such, we evaluate only performance with and without the extended instructions.

4.2   Supervised classification

To investigate the effect of domain adaptation on radical content detection, we consider two supervised classification tasks from the Counter dataset. Firstly, `Call for Action` assesses whether the content indicates a need for intervention, using a five-point scale ranging from *no action needed* to *urgent action*. Secondly, `Radicalization Level` which captures the degree of radicalization expressed in the content. It includes six levels: *Neutral*, *Expression of Radical Views*, *Use of Radical Propaganda*, *Affiliation with Radical Groups*, *Dehumanization of the Other*, and *Call for Action against others*. For both tasks, we use the contractor annotations included in the Counter dataset. We experiment with the other post-level annotations, e.g. named entities and ideological labels, as auxiliary tasks. We adopt the same 70:20:10 data split as Riabi et al. (2025). Each language subset, French and English, includes 2650 annotated examples. The distribution of examples across the dif-

ferent categories (ideology, source, classes) varies between the two languages, with noticeable imbalances in class representation. Matching the original work, we report the macro-F1 score.

Models We experiment with `xlmt` (Barbieri et al., 2022), an `xlmr` (Conneau et al., 2020) model which has been fine-tuned for the Masked Language Modelling (MLM) task on multilingual social media data. This model was used in the work of Riabi et al. (2025) to establish a benchmark for the `Call for Action` task. As an encoder-only model, its architecture is more appropriate for the supervised classification task compared to the instruction-tuned decoder-only models used in Sec. 4.1 (Yu et al., 2023; Edwards and Camacho-Collados, 2024). We use the MaChAmp v0.2 toolkit (van der Goot et al., 2021), which supports a multi-task setting with a shared encoder and separate classification heads for each task. These heads are jointly fine-tuned during training, and the encoder parameters are updated based on the combined gradients from all tasks. We report hyperparameters in App. D.2.

Domain adaptation We aim to adapt the base `xlmt` encoder to the language characteristics specific to these online platforms and assess how this adaptation impacts downstream tasks within the Counter dataset. We use the IYKYK dataset to continue training the encoder model using the MLM objective, before fine-tuning it for classification as described above. After filtering out one-word or empty posts, we include approximately 9M posts from each forum. We experiment with three settings: fine-tuning on each forum separately and fine-tuning on the combined posts from both forums. In each case, we split the data into 90% for training and 10% for validation. Hyperparameters are reported in App. D.2.

5   Lexicon induction pipeline

The tasks in Sec. 4.1 require confirmed positive and negative examples of in-group language for each community. While existing lexicons provide a basis for positive examples, these lexicons tend to rely on out-of-date data or are limited in size. To construct a gold standard reference of in-group language for each community, we follow the approach of de Kock (2025), which entails combination of statistical methods and human expert labelling.

Candidate generation In-group language constitutes a small fraction of the general vocabulary

of a group; as such, automated methods have been developed to extract it. We use the `LISTN-C` method (de Kock, 2025), which allows us to identify words that, based on their usage patterns, have a higher likelihood of being particularly relevant to the group. We extract the top 3050 words per community, using a lower threshold to ensure that it will include both positive and negative test cases. The system is trained separately for each community, using annual data snapshots from 2018 to 2024 from the IYKYK dataset. Further implementation details are provided in App. B.

Expert annotation   Given a set of candidate words for each community and 10 usage examples[4] for each word, human experts are employed to assess whether the word constitutes valid in-group language for the community in question. Drake (1980) states that a primary purpose of in-group language is to signal cohesion and solidarity with a group. As such, our expert labelling is based on the question: *If you saw someone use this word in a neutral setting, would it increase your expectation that they belong to the relevant community?* This would be strongly indicative, but is not a strict requirement for inclusion. Our full annotation instructions, including examples of positives and negatives, are provided in App. B. The annotators are doctoral scholars in sociology who specialise in the online rhetoric of these groups. They are employed as research assistants in this project through our institution, and receive a salary of $37/hour. A pilot annotation round is performed by one author and both annotators using 300 samples per platform, yielding a Cohen's Kappa of 0.813 and 0.718, respectively. The latter value is for Stormfront, which has a much larger class imbalance (discussed below), hence the lower agreement score is not unexpected.

Gold standard in-group language   The full candidate sets are annotated by one expert annotator each, resulting in 1401 positive test cases for Incels and 399 for Stormfront. The remainder of the 3050 samples for each community are confirmed negatives. The discrepancy in the number of positive instances per community is substantial, but not unexpected: Bogetić (2023) states that the incels are more linguistically innovative compared to the alt-right. Furthermore, we note that the Incels vocabulary is inflated due to a set of morphemes that

---

[4]The examples coincide with those used in Sec. 4.1.

chad: *chadstralian, chadlike, chadland, chadsplaining, chadsexuality, subchads, chadspawn, chadstein, chadless, chadface, chadfisher, chadder, chadstruck, chaddier, chadworship, terrachad, uberchad ...*

(f)oid: *australoid, schizoids, roastoid, rightoids, noodlefoid, angloids, caucasoids, normoid, mayofoids, westoids, leftoid, swarthoid, foidish, insectoid, brownoid ...*

Table 2: Variants of popular subwords on Incels.

are repeatedly used to derive new, somewhat obscure compound words, some of which are shown in Table 2. Of the 1401 positive terms for Incels, 570 include the substrings *pill*, *mog*, *maxx*, *cel*, *chad*, *cuck*, and *oid*. This has been observed in prior work: Yoder et al. (2023) detect over 1500 unique words that include the suffix *-cel*, used to refer to an incel with some characteristic (e.g., *tallcel* might refer to an incel who is tall).

6   Results for zero-shot evaluation

6.1   Task 1: Classification

For both platforms, the highest F1 score (64.4% for Stormfront and 80.0% for Incels) is achieved by `Llama-3.3` (70B). Providing full `instructions` leads to increases in F1 of 9.5% and 10.0%, respectively, over the `definition`-only framing. Adding additional context (i.e., example posts) does not consistently lead to further increases in F1, but it does result in higher recall (shown in Fig. 6 in App. D) for most models and both platforms, while precision (Fig. 5 in App. D) suffers. In other words: including the post examples in the prompt increases the probability that the model will assign a positive label. This may be caused by the fact that the prompt indicates that the posts were sourced from the platform in question, creating a link to the group. As expected, the larger model variants tend to perform better than their smaller counterparts.

Matching model and context settings, the F1 scores are higher for Incels (left) than for Stormfront (right). This is likely impacted by the larger imbalance in the test set, also reflected in the random baseline. The commonly reused morphemes by the Incels (as discussed shown in Table 2) also impacts performance, shown in the middle panel of Fig. 3. These words are considered to be easier to identify based on their surface-level characteristics. Excluding the 570 positive terms that contain *pill*, *mog*, *maxx*, *cel*, *chad*, *cuck*, or *oid*, we observe an average drop of 13.5% in performance, which results in more similar scores in the best model for Incels (71.0%) and Stormfront (64.4%).
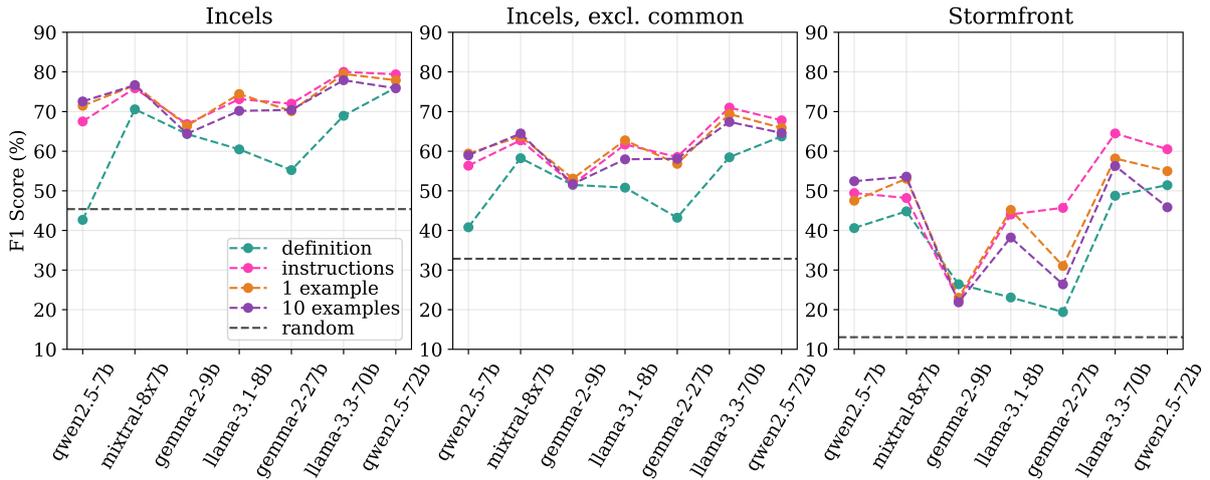
Figure 3: Results for Task 1 (Classification).

6.2 Task 2: Retrieval

Results for the retrieval task are shown in Fig. 7 in App. D. We observe very similar trends compared to the classification formulation in Task 1: a consistent improvement in F1 with the full instructions, a drop in F1 when excluding the common subword cases, and higher scores for Incels compared to Stormfront. The absolute values are similar as well, with an average difference of 8.6% and 7.4% observed for the `10-examples` setting for Incels and Stormfront, respectively. This is a somewhat surprising result given the prior expectation that this is a more challenging task, although the same objective is evaluated in both. We observe that the labels coincide with the Task 1 results in 71% (Incels) and 79% (Stormfront) of cases.
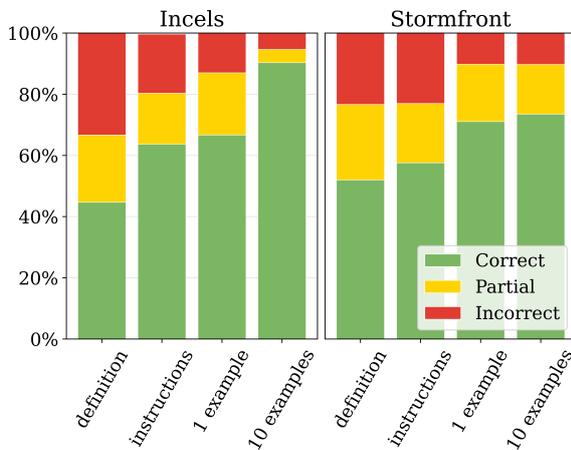
6.3 Task 3: Decoding

Figure 4: Results for Task 3 (Decoding).

For the decoding task (Fig. 4) we note a diver-

gence from the results in Task 1 and 2, in that a substantial improvement is achieved over the `instructions` framing by including post examples. For Incels, a minor improvement is gained from `1-example` and a substantial increase from `10-examples`, generating correct definitions for 90.3% of test cases, compared to 44.7% in the `definition` framing. For Stormfront, a large improvement is observed when adding a single example post, but the `10-examples` framing does not provide a substantial further increase, generating correct definitions in 73.5% of cases. We hypothesise that this is because the Stormfront language seems to be more heavily codified, often relying on acronyms or symbols. Considering the examples in Fig. 1, for example, the meaning of *KLASP* could not be inferred from context alone as an acronym for *Klannish Loyalty a Sacred Principle*[5]. They also tend to use more symbolism; for instance, the use of triple parentheses is a well-known antisemitic dogwhistle (Mendelsohn et al., 2023). By comparison, the Incels' language seems more focused on signalling in-group status rather than concealing its meaning from outsiders.

For both platforms, polysemous words are often incorrectly defined in lower-context framings. For example, *fob* is an derogatory term for immigrants, but can also refer to a key fob. All prompts prime the model to produce the in-group sense, so these cases should still be viable. Similarly, providing only one post can result in the model "overfitting" to that use case, sometimes performing worse than without it. For the term *apache*, a transphobic

---

[5] https://www.adl.org/resources/hate-symbol/klasp

meme[6], all 10 examples refer to literal helicopters. Including more posts may mitigate this issue, but it is also an inherent challenge of this task. In a small number of cases, models respond with a variant of: *I can't help with that request. The term provided contains hate speech and is not suitable for discussion.* This primarily occurs for the `definition` framing, which does not include the research motivation provided in the other framings.

For this task, we do not observe a large difference in the results for the common subword cases for Incels (shown in Fig. 8 in App. D). This indicates that while the surface forms may be easier to identify, their meanings are not necessarily straightforward. For example, variants of *-cel* are sometimes misinterpreted as referring to excelling.

7 Results for supervised classification

Results for the supervised classification evaluation are shown in Table 3. We report F1 scores, averaged over 5 runs, for four different configurations for our two target tasks in English. In each case, the top row represents the baseline, i.e. the model of Riabi et al. (2024). The columns represent different fine-tuning configurations, where each combining the target task with an auxiliary task. Statistical significance is determined using a one-sided t-test.

The results demonstrate that domain adaptation on the IYKYK data significantly improves performance over the baseline model in most configurations for both tasks. We achieve state-of-the-art results on the `Call for Action` task, improving over previously reported benchmarks. Interestingly, the combined forum data setting (denoted xlmt$_{+strf+inc}$) frequently provides robust improvements beyond either of the platforms individually, despite the annotated data representing primarily Jihadist and alt-right ideologies. This suggests that merging data from diverse extremist sources enriches the models' capability to capture nuanced linguistic features common across radical communities.

The combination of domain adaptation with auxiliary tasks through multi-task learning consistently achieves the best performance. Integrating ideology prediction alongside domain adaptation enhances the macro-F1 scores by up to 5.2% for `Radicalization Level` detection.

We report the first results for the `Radicalization Level` prediction task, es-

---

[6]https://en.wikipedia.org/wiki/I_Sexually_Identify_as_an_Attack_Helicopter

Model	CfA	+ideology	+NER	+rad.level
xlmt	64.6	65.1	64.6	64.8
xlmt$_{+inc}$	68.1† (+3.5)	66.5 (+1.4)	67.3† (+2.7)	67.4† (+2.6)
xlmt$_{+strf}$	63.8 (−0.9)	68.3‡ (+3.2)	66.8† (+2.2)	67.7† (+2.9)
xlmt$_{+strf+inc}$	66.7† (+2.0)	68.5‡ (+3.4)	69.0‡ (+4.4)	66.6 (+1.7)

(a) Call for Action (CfA) prediction

Model	Rad.level	+ideology	+NER	+CfA
xlmt	56.4	56.9	58.3	56.5
xlmt$_{+inc}$	59.6 (+3.2)	62.1‡ (+5.2)	58.9 (+0.6)	60.3‡ (+3.8)
xlmt$_{+strf}$	59.6† (+3.2)	59.1† (+2.1)	59.7 (+1.5)	60.6‡ (+4.2)
xlmt$_{+strf+inc}$	60.0† (+3.6)	61.3‡ (+4.3)	60.5 (+2.2)	60.1† (+3.6)

(b) Radicalization Level prediction

Table 3: Effect of domain adaptation on radical content classification in English. We use a one-sided t-test to determine statistical significance of improvements relative to the baseline, with $\dagger$ indicating $P < 0.05$ and $\ddagger$ indicating $P < 0.01$.

tablishing a baseline for future work. Predicting these labels is considered more challenging due to definitional ambiguity and contextual sensitivity, which is reflected in the comparatively lower performance scores observed in our experiments.

Table 7a in App. D.2 shows the results for French. Compared to English, domain adaptation yields smaller gains. As the IYKYK data is primarily in English, the limited impact on French performance is expected.

Detecting hate speech A central challenge for toxic language detection is the non-transferability of models across datasets due to labelling inconsistencies (Fortuna et al., 2021; Khondaker et al., 2023) and vocabulary differences (Röttger et al., 2021). Such concerns create particular challenges in applying toxicity and hatespeech detection technologies to communities which have high degrees of in-group language and have rapid linguistic evolution. In Table 8 in App. D, we test if hate speech detection models benefit from domain adaptation on the IYKYK data. We fine-tune the domain adapted models for classification on the dataset of Montariol et al. (2022), using the same procedure as for radical content detection. Matching our expectations, performance on anti-immigrant hate speech increases when training on Stormfront, and performance on misogynist hate speech increases when training on Incels. This suggests that the insights in this paper may apply to toxic language detection problems more broadly.

8 Discussion

The results in Sec. 6 and 7 indicate that it is necessary to first adapt NLP tools before application on extremist cryptolects. For instruction-tuned models, this can be achieved by providing context details in the prompt. For encoder models, domain adaptation on data from extremist platforms significantly improves performance. These results have significant implications for automated content moderation systems, which often leverage LLMs[7].

Toxic language is commonly removed during pre-training of NLP models to prevent the generation of such content (Rae et al., 2022; Anil et al., 2023; Longpre et al., 2024; Soldaini et al., 2024). A consequence of such removals is that such tools are ill-suited for tasks relating to toxicity (e.g. modelling extremism) due to shifts in the underlying data distribution. Indeed, the removal of toxic content also has implications for the representation of other communities, such as queer communities in LLMs (Xu et al., 2021). Thus, two competing goals arise: on one hand, toxic language must not be generated by NLP models; on the other, the removal of toxic language during pre-training can limit the applicability of models to settings where good representations of toxicity and related concepts are paramount. This suggests a need for specialised models for generating content and for monitoring radical content and hate speech, respectively. The resources developed in this work provide a useful starting point for future work to develop such tools.

9 Conclusion

We conduct the first large-scale investigation of LLM performance on extremist cryptolects. Evaluating eight models over six tasks and two communities, our results provide useful insights for the use of LLMs in this domain. In particular, we find that the performance of instruction-tuned LLMs can be substantially improved by introducing additional context in the prompt. Furthermore, domain adaptation of encoder models on data from extremist platforms improves the detection of radical and toxic content. This study contributes to a limited but growing body of work in NLP on the topic of extremist language. We hope that the resources created in this work, including large datasets and expert-validated lexicons, will support further work in this important area.

---

[7]E.g., a recommended usage of the Gemini family of models in the official documentation.

Ethics

The ethical implications of research on radicalization require careful consideration. The research conducted in this work has been approved by an independent Ethics Review Board within our institution. It was considered to be a low-risk application since all data used in this work is publicly available (without registration gating) and users interact on an anonymous basis. Both platforms' Terms and Conditions explicitly state that all posts are openly accessible and that users should consider this in their disclosures. While we cannot guarantee that users have not inadvertently provided personally identifiable information in their posts, the platforms that we have considered in this work discourage and block personal information where it is detected. In our discussion of the platforms' content in this paper, we do not include usernames or other personal details.

We consider it important to ensure these datasets are accessible for research while avoiding unintended harms. While the models and datasets we develop can be used to support intervention and moderation to enhance social welfare, there is a risk of misuse, including the generation of toxic content and unjust profiling or preemptive prosecution. As such, we only make our models and datasets available to verified researchers upon request and under a memorandum of understanding requiring responsible use of the resources. Predictive models must be used carefully within broader intelligence systems, ensuring transparency and incorporating socio-cultural contexts to promote fair and responsible use (Winter et al., 2021).

Handling harmful text, such as hateful or radical content, poses challenges for researchers, annotators, and dataset curators. Repeated exposure to violent or hateful material can cause psychological harm, including vicarious trauma, highlighting the need for safeguards to protect those working with this data (Kirk et al., 2022; Riabi et al., 2024). In our evaluations, we only work with annotators for whom researcher risk management plans are in place. This includes consideration for mental health and personal safety aspects. In selecting examples for this paper, care was taken to ensure that the content is not too distressing to read, noting that there is a substantial psychological burden on researchers who work in this domain.

Limitations

Since the trends we observe in our results are consistent over multiple communities, datasets and tasks, we believe our insights to be sound. However, we recognise that defining in-group language, like defining extremism, is subjective in nature. For this reason, we develop and share detailed annotation instructions, and perform pilot annotations to ensure alignment. We also prioritise high quality annotations by PhD-level experts. Even so, using a single annotator per dataset is a limitation; ideally, multiple annotations would be obtained for every sample.

In this work, we do not account for the effect of time. We include posts from 2018 to 2024 in our analysis of in-group language. While our annotators have years of experience working on these groups, we recognise that the language in these groups is very dynamic and that the annotators may have a bias towards more recent language, which may play a role in their interpretations.

Finally, many different extremist ideologies exist, and it is likely that the insights from this work would not fully apply to all of them. In fact, our results point to the fact that there are large differences in the performance for the two groups evaluated here. By including two communities in our study, we aim to provide a balanced impression and to focus on trends rather than on absolute values when interpreting our results. We hope that future work will use the framework proposed here to explore other ideologies.

Acknowledgments

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, and 109 others. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Stephane J Baele, Lewys Brace, and Travis G Coan. 2021. From "incel" to "saint": Analyzing the violent worldview behind the 2018 toronto attack. *Terrorism and political violence*, 33(8):1667–1691.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *lrec:2022:1*, pages 258–266, Marseille, France. European Language Resources Association.

Omran Berjawi, Giuseppe Fenza, and Vincenzo Loia. 2023. A comprehensive survey of detection and prevention approaches for online radicalization: Identifying gaps and future directions. *IEEE Access*, 11:120463–120491.

Ksenija Bogetić. 2023. Race and the language of incels: Figurative neologisms in an emerging english cryptolect. *English Today*, 39(2):89–99.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *acl:2020:main*, pages 8440–8451, Online. acl.

Maura Conway. 2006. Terrorism and the internet: New media—new threat? *Parliamentary Affairs*, 59(2):283–298.

Christine de Kock. 2025. Inducing lexicons of in-group language with socio-temporal context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13281–13291, Vienna, Austria. Association for Computational Linguistics.

Christine de Kock. 2026. Jointly modelling the evolution of social structure and language in online communities. In *Proceedings of the International AAAI conference on web and social media*.

Christine De Kock and Eduard Hovy. 2024. Investigating radicalisation indicators in online extremist communities. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 1–12.

Donatella Della Porta and Gary LaFree. 2012. Processes of radicalization and de-radicalization. *International Journal of Conflict and Violence (IJCV)*, 6(1):4–10.

Glendon Frank Drake. 1980. The social role of slang. In *Language*, pages 63–70. Elsevier.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia. ELRA and ICCL.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

M. Gaikwad, Swati Ahirrao, Shraddha Phansalkar, and K. Kotecha. 2021. Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access*, 9:48364–48404.

Mayur Gaikwad, Swati Ahirrao, Shraddha Phansalkar, Ketan Kotecha, Shalli Rani, and Lorenzo Putzu. 2023. Multi-ideology, multiclass online extremism dataset, and its evaluation using machine learning. *Computational Intelligence and Neuroscience*, 2023.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ian Hancock. 1986. The cryptolectal speech of the american roads: Traveler cant and american angloromani. *American Speech*, 61(3):206–220.

DS Hoffman. 1997. Web of hate: Extremists exploit the internet. www.ojp.gov/ncjrs/. Accessed 18-04-2025.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Md Tawkat Islam Khondaker, Muhammad Abdul-mageed, and Laks Lakshmanan, V.s. 2023. Cross-platform and cross-domain abusive language detection with supervised contrastive learning. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 96–112, Toronto, Canada. Association for Computational Linguistics.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jasmine Latimore and John Coyne. 2023. Incels in australia: the ideology, the threat, and a way forward.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6929–6947, Toronto, Canada. Association for Computational Linguistics.

Elisa Mattiello and 1 others. 2005. The pervasiveness of slang in standard and non-standard english. *Mots Palabras Words*, 6:7–41.

Gretchen McCulloch. 2019. *Because internet: Understanding how language is changing*. Random House.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180.

Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363, Online only. Association for Computational Linguistics.

PMO. 2025. Downing street opens doors to adolescence creators for vital discussion on protecting our children. www.gov.uk/. Accessed 18-04-2025.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, and 61 others. 2022. Scaling language models: Methods, analysis & insights from training gopher. *Preprint*, arXiv:2112.11446.

Arij Riabi, Menel Mahamdi, Virginie Mouilleron, and Djamé Seddah. 2024. Cloaked classifiers: Pseudonymization strategies on sensitive classification tasks. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 123–136, Bangkok, Thailand. Association for Computational Linguistics.

Arij Riabi, Virginie Mouilleron, Menel Mahamdi, Wissam Antoun, and Djamé Seddah. 2025. Beyond

dataset creation: Critical view of annotation variation and bias probing of a dataset for online radical content detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8640–8663.

Ben Rich and Eva Bujalka. 2023. The draw of the 'manosphere': understanding andrew tate's appeal to lost men. The Conversation.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Aaron Shaw. 2023. Social media, extremism, and radicalization. *Science advances*, 9(35):eadk2031.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1683–1701, Mexico City, Mexico. Association for Computational Linguistics.

Daniele Valentini. 2021. Digital mind invasion: Far-right individuals and socio-technical disgust. *CYBERPOLITICS: Political Philosophy of the Future*, pages 125–143.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Gabriel Weimann and Ari Ben Am. 2020. Digital dog whistles: The new online language of extremism. *International Journal of Security Studies*, 2(1):4.

Charlie Winter, Peter Neumann, Alexander Meleagrou-Hitchens, Magnus Ranstorp, Lorenzo Vidino, and

Johanna Fürst. 2021. Online extremism: Research trends in internet activism, radicalization, and counter-strategies. *International Journal of Conflict and Violence*, 14:1–20.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Michael Yoder, Chloe Perry, David Brown, Kathleen M Carley, and Meredith Pruden. 2023. Identity construction in a misogynist incels forum. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 1–13.

Hao Yu, Zachary Yang, Kellin Pelrine, Jean-François Godbout, and Reihaneh Rabbany. 2023. Open, closed, or small language models for text classification? *CoRR*, abs/2308.10092.

A Hyperparameters

A.1 Lexicon induction

The LISTN-C method consists of *(i)* jointly learning temporal user and word embeddings in a shared space, *(ii)* clustering users into sub-groups to account for social fragmentation, and *(iii)* extracting the maximum relevance words based on their embedding distance to the community centroid(s). We use the optimal configuration from de Kock (2025), which assumes 5 clusters. While the original system investigates shorter-term linguistic change and therefore uses monthly data increments, we are interested in more stable terminology and therefore use annual snapshots. The original work studies lexical variation and therefore excludes standard words. Since we are also interested in dogwhistles, we maintain standard and non-standard words in our candidate set. We include only users active in more than one timestep and words used by more than 20 users. We exclude non-alphabetic characters. Hyperparameters are provided in Table 4. Complete details of the system are provided in de Kock (2026).

A.1.1 Domain Adaptation

Hyperparameters for the domain adaptation training are provided in Table 5.

A.1.2 Supervised classification

Hyperparameters for the supervised classification task are provided in Table 6.

Hyperparameter	Value
K	100
c_0	0.005
λ_1	1
λ_2	10
Learning rate	1e-05
Batch size	40
Seed	42
Maximum epochs	10000
Optimizer	AdamW
Validation	0.05
Early stopping	P=100, tolerance=0.001

Table 4: Hyperparameters used for the LISTN-C system. Models are trained on a single NVIDIA GeForce GTX 108 GPU.

Hyperparameter	Value
Learning rate	8e-05
Batch size	32
Seed	42
Distributed training type	Multi-GPU
Number of devices	8
Gradient accumulation steps	8
Total training batch size	2048
Number of epochs	3
Optimizer	AdamW
Learning rate scheduler	Linear
Warmup ratio	0.01

Table 5: Hyperparameters used for MLM training. We used the final checkpoint for all three models, selected based on performance on the development set. We conduct our experiments on RTX8000 GPUs.

B Annotation instructions

Annotation instructions for identifying and defining in-group language are provided in this section.

B.1 Identifying in-group language

B.1.1 Incels

The goal of this work is to investigate in-group language within online extremist groups. In-group language is defined as specialised language developed by a social group to (1) indicate closeness to the in-group, and (2) create distance from the out-group. One way to think about it is: if you saw someone use this word in a neutral setting, e.g. Twitter, would it increase your expectation that they belong to an incel community? This would be strongly indicative, but is not a strict requirement for inclusion.

You will be given a list of words and a number of examples of these words being used on incels.is. These words have been identified as potential examples of in-group language to be used in a downstream analysis, which will compare the language

Hyperparameter	Value
Encoder dropout	0.2
Update encoder weights	True
Batch size	8
Learning rate	1e-4
Weight decay	0.01
Optimizer betas	(0.9, 0.99)
Learning rate scheduler	Linear decay with warmup
Gradual unfreezing	True
Discriminative fine-tuning	True
Number of epochs	10

Table 6: Fine-tuning hyperparameters. The best checkpoint for each model is selected based on its performance on the development set. We conduct our experiments on RTX8000 GPUs.

in this community with a different group. It is expected that there will be invalid words in this list. We would like you to help us to filter them out. You don't need to read all the examples; they are only provided as context.

Positive examples include:

- *Lexical innovations (non-standard words) invented by this group, e.g. foids. Standard words can be thought of as what you might read in Wikipedia or a news source.*

- *Dogwhistles (standard words that have taken on a specific meaning to the group that is not held by the broader population), e.g. Becky, cope*

- *Linguistic exports, or words that originated in the incel community but are now used more broadly, e.g. incel, mogging*

- *Linguistic imports, or words that originated in another space and is elsewhere with the same meaning, but now belongs to the specialised language of this group, e.g., cuck, slut, hyoid, clavicles, gynocentric, alpha*

- *Names of people significant to the community*

A word can belong to more than one of these categories.

Invalid words include:

- *Standard language used in a standard way, e.g. internet*

- *Internet language that is not specialised to this group, e.g. becuz*

- *Spelling errors that are not specialised to this group, e.g. litterally.*

- *Non-English text - there is a bit of this; please mark it as invalid.*

- *Names of people or places, unless they are specifically relevant to this group. For example: Andrew Tate / Elliot Rodger = Yes, Anthony Albanese = Probably not.*

- *Usernames (should be apparent from context).*

This is a subjective task, and ultimately we are relying on your expert knowledge of this domain to make a judgment. Having said that, note that the idea is that we want to filter out words that are definitely invalid, so in the case of borderline words, you should opt to include them.

B.1.2 Stormfront

The goal of this work is to investigate in-group language within online extremist groups. In-group language is defined as specialised language developed by a social group to (1) indicate closeness to the in-group, and (2) create distance from the out-group. One way to think about it is: if you saw someone use this word in a neutral setting, e.g. Twitter, would it increase your expectation that they belong to an alt-right community? This would be strongly indicative, but is not a strict requirement for inclusion.

You will be given a list of words and a number of examples of these words being used on Stormfront.org. These words have been identified as potential examples of in-group language to be used in a downstream analysis, which will compare the language in this community with a different group. It is expected that there will be invalid words in this list. We would like you to help us to filter them out. You don't need to read all the examples; they are only provided as context.

Positive examples include:

- *Lexical innovations (non-standard words) invented by this group, e.g. foids. Standard words can be thought of as what you might read in Wikipedia or a news source.*

- *Dogwhistles (standard words that have taken on a specific meaning to the group that is not held by the broader population), e.g. Becky, cope*

- *Linguistic exports, or words that originated in the alt-right community but are now used more broadly, e.g. incel, mogging*

- *Linguistic imports, or words that originated in another space and is elsewhere with the same meaning, but now belongs to the specialised language of this group, e.g., cuck, slut, hyoid, clavicles, gynocentric, alpha*

- *Names of people significant to the community*

A word can belong to more than one of these categories.

Invalid words include:

- *Standard language used in a standard way, e.g. internet*

- *Internet language that is not specialised to this group, e.g. becuz*

- *Spelling errors that are not specialised to this group, e.g. litterally.*

- *Non-English text - there is a bit of this; please mark it as invalid.*

- *Names of people or places, unless they are specifically relevant to this group. For example: Andrew Tate / Elliot Rodger = Yes, Anthony Albanese = Probably not.*

- *Usernames (should be apparent from context).*

This is a subjective task, and ultimately we are relying on your expert knowledge of this domain to make a judgment. Having said that, note that the idea is that we want to filter out words that are definitely invalid, so in the case of borderline words, you should opt to include them.

B.2 Defining in-group language

Given a word and a set of definitions, assess the validity of each definition in the context of [Incels.is\Stormfront.org].

These definitions are generated by a large language model. The goal is to investigate the quality of the definitions generated using different approaches. The definitions for each word are randomly ordered and does not correspond to the same approach every time.

NOTE: More than one definition can be correct (or incorrect, or partially correct etc) for each word. There does not need to be a "best" one.

Possible answers are:

- *Accurate and complete: a valid definition that captures all important aspects of the term.*

- *Partially correct: Correct in some ways but missing some nuance. Note: it is preferable not to use this option if one of the others might apply.*

- *Inaccurate: Does not capture the meaning of the word in this context.*
- *No Answer: Model refuses to answer or provides a nonsensical output, e.g. ""I cannot help you with that."" or ""<your definition>""*

C Prompts

For each model, we use the appropriate prompt format as per the model card and a temperature of 0.3. Tests are run through `together.ai` for a total of $154 in API costs. Outputs are parsed by extracting the text following *ANSWER*. In addition to the instructions below, we also provide the required output format and demarcations of the instructions and task.

C.1 Task 1: Classification

C.1.1 Definition

In-group language is defined as specialised language developed by a social group to (1) indicate closeness to the in-group, and (2) create distance from the out-group. You will be given a word. Your job is to determine if it is in-group language.

C.1.2 Instructions, 1-example and 10-examples

The prompt is constructed based on the instructions in Sec. B.2, preceded by: *You are an expert sociologist working on online extremist groups.* Then, we provide either no examples (referred to as *instructions*), a single post (*1-example*) or a list of ten (newline-separated) posts (*10-examples*).

C.2 Task 2: Retrieval

C.2.1 Definition

In-group language is defined as specialised language developed by a social group to (1) indicate closeness to the in-group, and (2) create distance from the out-group. You will be given a post from an online forum. Your task is to identify instances of in-group language in this post. Return a list of words that are positive examples according to the description of in-group language provided. If there are none, return an empty list.

C.2.2 Instructions

The prompt for this task consists of the instructions in Sec. B, preceded by: *You are an expert sociologist working on online extremist groups.* This instructions are followed by: *You will be given a post from [Stormfront.org | Incels.is]. Your task is to identify instances of in-group language in this post. Return a list of words that are positive examples according to the description of in-group language provided. If there are none, return an empty list.*

C.3 Task 3: Decoding

C.3.1 Definition

The goal of this work is to investigate in-group language within online groups. You will be given a word used in an online conversation. This word has been identified as an example of in-group language, which includes non-standard language. Your task is to infer the meaning of the word.

C.3.2 Instructions, 1-example, and 10-examples

The prompt is constructed based on the following description, followed by one, ten or zero examples: *The goal of this work is to investigate in-group language within online extremist groups. You will be given a word used on [Stormfront.org, an online alt-right community | Incels.is, an online anti-women community]. This word has been identified as an example of in-group language, which includes non-standard language. Your task is to infer the meaning of the word.*

D Additional results

D.1 Zero-shot evaluation

Precision and recall scores for Task 1 are shown in Figures 5 and 6, respectively. Fig. 7 contains the F1 scores for Task 2. The results for the common subwords variant for Incels for Task 3 are shown in Fig. 8.

D.2 Detecting radical content

Results for the radical content detection task in French are provided in Tables 7a and 7. Results for the hatespeech detection task are in Table 8.
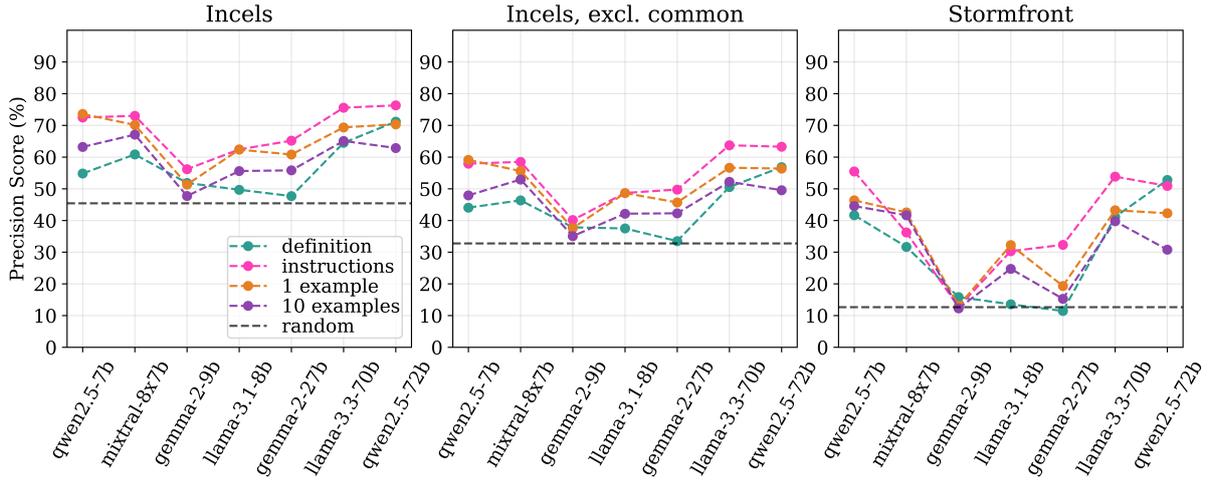
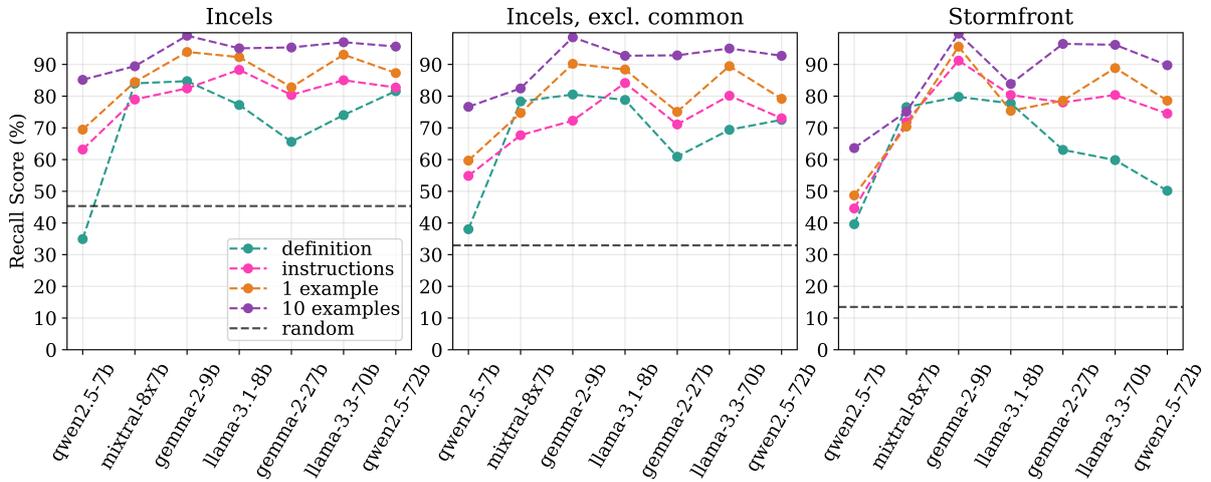Figure 5: Precision scores for Task 1.

Figure 6: Recall scores for Task 1.

Model	CfA	CfA+ideology	CfA+NER	CfA+rad. level
xlmt	65.7	65.6	62.0	63.9
xlmt$_{+incels}$	65.0 (−0.6)	68.7 (+3.2)	63.8 (+1.9)	64.2 (+0.3)
xlmt$_{+stormfront}$	65.9 (+0.3)	62.3 (−3.3)	61.7 (−0.3)	62.5 (−1.4)
xlmt$_{+stormfront+incels}$	65.8 (+0.2)	64.0 (−1.6)	63.8 (+1.8)	59.4 (−4.5)

(a) Call for Action (CfA) prediction

Model	Rad. level	Rad. level+ideology	Rad. level+NER	Rad. level+CfA
xlmt	61.7	59.8	56.9	59.1
xlmt$_{+incels}$	60.0 (−1.7)	60.0 (+0.2)	57.7 (+0.8)	58.4 (−0.7)
xlmt$_{+stormfront}$	57.8 (−3.9)	60.2 (+0.4)	57.3 (+0.4)	58.1 (−1.0)
xlmt$_{+stormfront+incels}$	58.8 (−2.9)	58.4 (−1.3)	57.6 (+0.8)	58.2 (−0.9)

(b) Radicalization Level prediction

Table 7: Effect of domain adaptation on radical content classification macro-F1 scores (%) for French, averaged over 5 runs. Columns represent different fine-tuning configurations, where each setting combines the target task (Call for Action or Radicalization Level detection) with an auxiliary task to assess its influence on performance. Main scores are shown in black, with the subscript indicating the delta relative to the xlmt baseline. Green subscripts denote an improvement over the baseline, red subscripts a drop. Statistical significance is determined using a one-sided t-test on the five run scores. A dagger (†) indicates $p < 0.05$, and a double-dagger (‡) indicates $p < 0.01$.

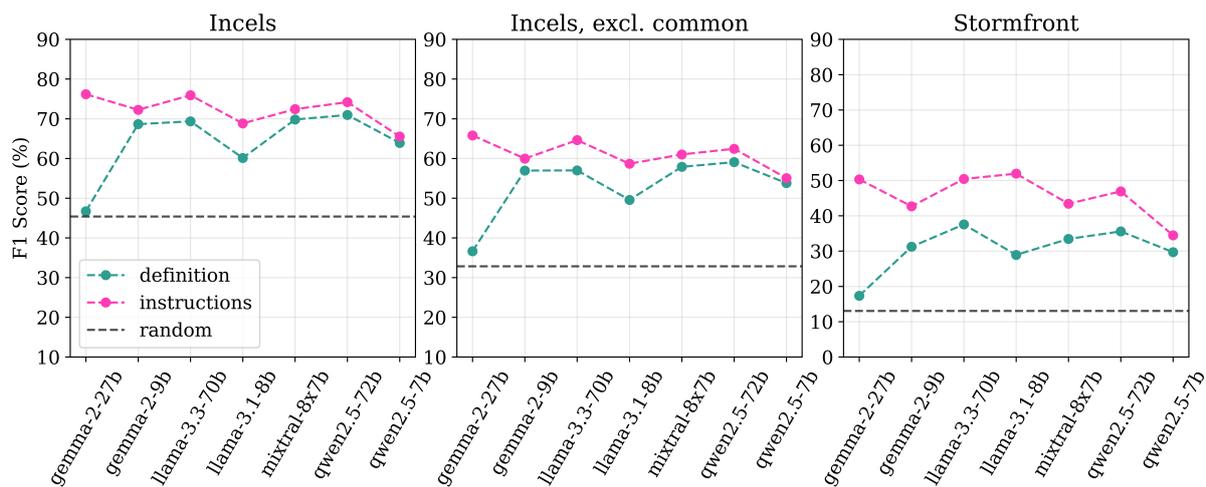Figure 7: Results for Task 2 (Retrieval).
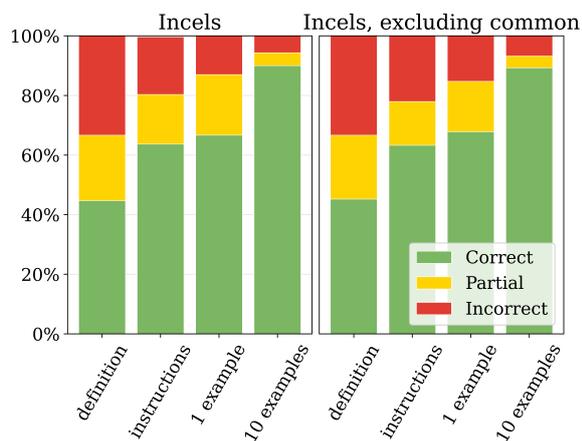
Figure 8: Results for Task 3, excluding common morphemes for Incels.

Model	immigrants	women
xlmt	77.0	78.8
xlmt$_{+incels}$	76.0 (−0.9)	79.9 (+1.1)
xlmt$_{+stormfront}$	78.1 (+1.1)	78.8 (−0.0)
xlmt$_{+stormfront+incels}$	76.9 (−0.0)	79.4 (+0.6)

Table 8: Hate speech detection macro-F1 scores on the English using corpora from HatEval and the splitting from Montariol et al. (2022). We use same experimental setting for supervised classification as radicalization detection.