An Evaluation of Language Models for Hyper-partisan Ideology Detection in Persian Twitter

Sahar Omidi Shayegan[1,3], Isar Nejadgholi[2], Kellin Pelrine[1,3], Hao Yu[1,3], Sacha Levy[1,3], Zachary Yang[1,3], Jean-François Godbout[3,4], Reihaneh Rabbany[1,3]

[1]McGill University, [2]National Research Council Canada, [3]Mila - Quebec AI Institute, [4]Université de Montréal

{sahar.omidishayegan, kellin.pelrine, hao.yu2 , sacha.levy, zachary.yang}@mail.mcgill.ca

isar.nejadgholi@nrc-cnrc.gc.ca, jean-francois.godbout@umontreal.ca, reihaneh.rabbany@mcgill.ca

Abstract

Large Language Models (LLMs) are now capable of successfully identifying the political beliefs of English-speaking social media users from their posts. However, assessing how LLMs perform in non-English languages remains difficult. In this work, we contribute to this area of research by determining the extent to which LLMs can predict the political ideologies of users on Persian social media. We begin by discussing the challenges associated with defining political parties within the Persian context and propose a solution based on a technique designed for the detection of hyper-partisan ideologies on social media. We create a new benchmark and show the potential and limitations of both open-source and commercial LLMs in classifying the hyper-partisan ideologies of users. We compare these models with smaller fine-tuned ones, both on the Persian language (ParsBERT) and translated data (RoBERTa), and confirm that they considerably outperform generative LLMs in this task. We further demonstrate that the performance of the generative LLMs degrades when classifying users based on their tweets instead of their bios, even if tweets are added as additional information; whereas the smaller fine-tuned models are more robust and achieve similar performance for all input settings. This study represents a first step toward political ideology detection in Persian social media, with implications for future research to understand the dynamics of political conflicts in Iran.

Keywords: Computational Social Science, Persian Language, Ideology Prediction

1. Introduction

Political ideology detection using Twitter data (now X) has been extensively studied in the English language (e.g., Pelrine et al. (2023); Yu et al. (2023); Törnberg (2023); Barberá (2015); Pennacchiotti and Popescu (2011)). The few studies that focus on other languages are generally limited to Western democracies, where the analysis of political campaigns and elections on social media has been used to monitor shifts in public opinion and the interactions between different ideological groups (Rodríguez-García et al., 2022; Chen et al., 2017; Jiang et al., 2022). Therefore, there is a significant research gap in studies conducted in other languages and those focusing on different types of political systems . In this work, we address this important limitation by focusing on the case of Iran. Despite the pivotal role that this platform has played in influencing political narratives in this country (Khorramrouz et al., 2023; Kermani and Tafreshi, 2023), it remains difficult to understand how political conflicts unfold between supporters and opponents of the Iranian regime.

The task of delineating the ideological orientation of supporters and opponents to the Islamic Republic of Iran poses several challenges. Indeed, unlike democratic countries where political parties are well-defined, the main division in Iran is largely driven by political ideology, which is not channeled through organized and institutionalized partisan groups (Azadi and Mesgaran, 2021). Thus, in the absence of distinct political parties, our research focuses on the more direct computational task of categorizing distinct ideological markers, specifically, hyper-partisan users representing two extreme viewpoints: "Pro-Government", the government supporters committed to the principles of the Islamic Republic; and "Pro-Monarchy", those who favour the return of the former monarchical regime. We recognize that there are several other political ideologies in the Iranian political space, including secularists, reformists, women's rights activists and Kurdish activists. However, for this study, we have decided to group all of these remaining ideologies under the class of "Others". We acknowledge this limitation and leave the more in-depth analysis of this last category for future research.

Our analysis of hyper-partisan ideology prediction in Persian Twitter focuses on data collected during the Woman-Life-Freedom movement, from October 18^{th} 2022 to January 11^{th} 2023. This period saw a significant surge in Persian tweets, with users extensively employing political and ideological hashtags and key terms. The importance of this event makes this time frame crucial for understanding the dynamics of political conflicts in Iran. We first labelled the users in our data by relying on clear ideological stances declared in the Twitter bios of users. We refer to these users as hyper-partisan users. This approach allowed us to anchor our research on users from contrasting ideological

2nd EURALI Workshop, pages 51–62
25 May, 2024.

backgrounds who are forthright about their beliefs, thereby ensuring minimal overlap of classes and mitigating the risk of mislabeling. Those without explicit indicators that failed to align with either one of these two groups were classified under an "Others" category, indicating a broader ideological spectrum. In this study, we specifically explored two tasks: 1) classifying hyper-partisan users based on the text in their bios and 2) classifying hyper-partisan users based on various combinations of the text found in their bios and in their tweets.

We, then, investigated the performance of different Large Language Models (LLMs) for identifying the above groups. Inspired by the widespread acclaim and proven efficacy of LLMs in diverse NLP tasks, including ideology prediction for English social media data (Tornberg, 2023; Yu et al., 2023), we evaluate these models for hyper-partisan ideology prediction within our labelled dataset. We begin with a comprehensive assessment of GPT-3.5, and then moved to other forms of LLMs, including open-source conversational LLMs such as Llama 2 Chat and WizardLM and smaller fine-tuned classifier models like RoBERTa (Liu et al., 2019a) and ParsBERT (Farahani et al., 2021).

Our results show that GPT-3.5 can classify bios with clear ideological markers reasonably well. However, this model is limited in the level of detail it can handle in the prompt and performs optimally only when all of its inputs are translated into English. Open-source conversational LLMs, such as WizardLM and Llama 2, achieve similar performances but also only when the data is translated into English. On the other hand, fine-tuned BERT-family LMs, both in Persian (ParsBERT) and English (RoBERTa), significantly outperform all generative LLMs. Overall, our results confirm that classifying tweets is a more challenging task for generative LLMs rather than classifying users based on the information found in their bios. More specifically, adding tweets to the input obfuscates the results of GPT-3.5, improves the classification performance of fine-tuned ParsBERT, and does not have a significant impact on fine-tuned RoBERTa. The main contributions of this paper are as follows:

- We evaluate various hyper-partisan ideology detection methods on Persian Twitter using different open and closed-source LLMs. Our work is a first step towards an area previously understudied despite Twitter's significance influence in Iran's political debates.
- Focusing on a period with a surge in Persian political tweets, we collect and label a new benchmark of Persian posts for this task and classify them into three main ideological groups: "Pro-Government", "Pro-Monarchy", and a third group, "Others", comprising various opposition factions.
- We present a comprehensive analysis of the potential and limitations of GPT-3.5 compared with other generative LLMs and fine-tuned classifiers. We also offer some insights into their efficacy in Persian and other low-resource language contexts.

2. Background and Related Work

This work is at the intersection of Persian NLP, and political ideology detection on social media. Here, we review the related work in each of these areas of research.

2.1. NLP in Persian

Although a large number of people speak Persian (there are approximately 110 million Persian speakers worldwide), very few language resources have so far been developed in this language. Shamsfard (2019) discuss the challenges of studying Persian and the reasons why it should be considered a low-resource language. They emphasize the need for effective solutions to leverage the potential of NLP techniques to create more resources for the automatic processing of Persian data.

There have also been efforts in creating foundational models in Persian, including ParsBERT (Farahani et al., 2021), GPT2-Persian (Khashei, 2021), ALBERT Persian (Farahani, 2020). Besides these general-purpose pre-trained Persian models, ARMAN has been specifically trained for text summarization in this language as well (Salemi et al., 2021). Furthermore, Persian is included in several multilingual pre-trained language models, including mBERT (Devlin et al., 2018), and XLMR (Conneau et al., 2020). Finally, Persian is also included in recently released generative AI models, such as LLaMA (Touvron et al., 2023a) and ChatGPT (Radford et al., 2021), but this language only represents a very small percentage of their training data. Indeed, while numerous studies have explored the application of generative LLMs in various of tasks beyond standard NLP benchmarks (Bandi et al., 2023; Ahuja et al., 2023; Weidinger et al., 2023; Bang et al., 2023), research such as Lai et al. (2023) and Zhu et al. (2023) has specifically evaluated the performance of the GPT-3.5 model in multilingual contexts, including Persian. However, their focus was not on political ideology detection on social media per se. Therefore, to the best of our knowledge, our study represents the first attempt to apply these more powerful models to this task in Persian.

2.2. Domain Background

Ideology Detection on Social Media: Ideology detection in online communities is a dynamic area of research that aims to classify and identify the partisanship or ideological leaning of of social media users (Pelrine et al., 2023; Pennacchiotti and Popescu, 2011; Chen et al., 2017). Thus far, most of this research has been focused on Twitter. For example, Yu et al. (2023) have examined how LLMs and smaller language models can be used to classify Twitter users according to their ideology. Their study involved three datasets, predominantly in English, related to the 2020 US election, the 2021 Canadian election, and COVID-19. They examined the capabilities of Llama 2, GPT-3.5, and RoBERTa, and found that RoBERTa outperformed the other two after fine-tuning. Additionally, they proposed to distinguish between "Explicit ideology" and "Implicit ideology". In this context, "Explicit" refers to classifying users based on their biographical information, which includes obvious ideological identifiers. On the other hand, "Implicit" involves predicting ideologies based on less explicit data, mainly a random set of users' tweets, which are less informative than the bio descriptions. Here, we employ a similar approach and consider classifying users based on their bios when they contain a strong indicator related to their tweets, which are sampled in different ways, as well as their combination.

Social Media and Ideology Analysis in Persian: We find several studies that have attempted to analyze social media activities on Persian Twitter. Notably, the work of Kermani and Tafreshi (2023) used the retweet graph, analyzed the political ideologies of Iranians during the 2017 presidential election and emphasized the significance of social media as a deliberative space for political discussions. Their results confirm that there were three communities active on Twitter during the election: reformists, conservatives, and diaspora. In another related work, Honari and Alinejad (2022), looked at bot activities on Twitter that supported controversial policies in Iran. Kermani and Hooman (2022) shed light on a significant feminist discourse among Iranian Twitter users during the summer of 2020. While the #MeToo movement emerged on this platform in Western countries in 2017, allowing millions of women to share their experiences of sexual abuse and harassment, Iranian users began discussing their own similar experiences on this platform three years later. The results of this study highlight the distinctions between Iranian feminism and its Western counterpart by highlighting the challenges of advocating for women's rights in Iranian society on social media.

Several studies also look at classifying Twitter users according to their ideological leanings, most notably to reveal the level of political change advocated by different political factions in Iran. For instance, Azadi and Mesgaran (2021) categorizes users into three distinct groups: "pro-regime", "dissidents", and "neutral individuals". Their work also focuses on two samples of Iranian Twitter users: the influencers and the ordinary people. They provide various statistical insights about these two samples, such as the age of their accounts, their time zone, and their interactions. They also classify some of the existing ideology clusters by focusing on their level of coordination and how much they are rooting for a regime change. In another work, Kermani (2023) confirm the extensive Twitter engagement of Iranian users in September 2022, despite all of the attempts by the government to impede online activism. Their analyses provide insights into the strategies used by pro-government agents to influence the debates and how the users overcame them. Finally, the work by Khorramrouz et al. (2023) examines the Mahsa Amini movement more specifically through the lens of gender equality. Their research reveals that the movement has intensified the polarization among Twitter users on this issue, with a more pronounced increase among those advocating for gender equality. Moreover, the authors categorize users into 'state-aligned' and 'pro-protest' groups, and argue that the pro-protest users align more closely with the baseline characteristics of Twitter users.

Overall, these studies help us identify the main ideological fault lines in the context of Iranian politics today. On the one hand, the main supporters of the government fall into the 'state-aligned' and 'pro-regime' categories of users. On the other hand, the dissidents encompass the 'monarchist', 'pro-protest', 'pro-women rights', and 'pro-minorities'. Since the other remaining dissidents users belong to a broad spectrum of (evolving and overlapping) ideologies without explicit markers, we have decided to group them in the "Others" category to minimize the risk of mislabeling. In this study, we only focus on the categories of "Pro-Government" and "Pro-Monarchy".

3. Dataset

Starting in September 2022, Persian Twitter users have been increasingly adding political hashtags to their tweets in response to political unrest in Iran. Using the Twitter Research API, we gathered our dataset by collecting real-time tweets between October 18^{th}, 2022, and January 11^{th}, 2023. Our data collection relied on a series of relevant political hashtags, which can be seen in Figure 1. We used 26 seed hashtags for this collection, both in Persian and English, which were identified by the authors who are familiar with the political context in Iran. A total of 231 million tweets were collected from 3.9 million users.

In the next step, the users were sorted by the

Language	Hashtags
Persian	مهسا_امینی، اعتصابات_سراسری، لبیک_یا_خامنه_ای، ایران_قوی
English	opiran, Mahsa_Amini, IranProtests2022

Figure 1: Examples of the hashtags used for crawling the tweets from Twitter.

Group	Keywords
Monarchy-supporters	👑، 🦁، پهلوی ، 🇭🇺
Government-supporters	سید علی ، شهادت ، #حاج_قاسم ، شهید ، ظهور

Figure 2: The indicator keywords used to find the most forthright supporters of groups.

number of times they were retweeted within the dataset, which ranks the more influential users first. Their Twitter biographical information was examined to find out if they were supporting one of the two extreme ideological views included in this study, “Pro-Government” and “Pro-Monarchy”.

We define the “Pro-Government” group as users who support the 1979 Islamic Revolution and the Islamic Republic of Iran—the current government in power. The “Pro-Monarchy” group are the users who support the Pahlavi dynasty and the former Imperial State of Iran.

We selected 1000 accounts (500 for each category) using a simple keyword search in their bio information in order to sample users who are likely to belong to either one of the categories of interest. Figure 2 shows some of these indicator keywords. We then ***labeled each user manually*** into three categories: *“Pro-Monarchy”*, *“Pro-Government”*, or *“Others”*. This led to a list of 382 “Pro-Monarchy” users, 316 “Pro-Government” users, and 302 users that could not be classified in those two opposing categories. Furthermore, we filtered out about 10% of the users who had excessively strong ideological keywords on their biographies since we considered them too easy for the classification task.

After this filtering, we are left with 909 users. We split the final dataset to train, validate, and test with the ratio of 0.4, 0.1, and 0.5, respectively. This resulted in 363 train samples, 91 validation samples, and 454 test samples. All the reported results are on the test set. The data collection flow is illustrated in Figure 3.

In Table 1, we included several examples of tweets posted by users classified in each group. To protect the privacy of the Twitter users, we do not include any biographical information in this paper. We also paraphrased and translated the example of tweets shown in Table 1. These messages show that the tweets supporting the monarchy evoke historical symbols and slogans associated with the pre-revolutionary era, such as the “lion and sun” flag and references to “Javid Shah” (Long Live the King), which is directly associated with the Pahlavi dynasty. The government-supportive tweets use language and imagery that reinforce loyalty to the Islamic Republic and its religious leadership, as seen in hashtags such as “#Labbaik_Ya_Khamenei” (I am at your service, Khamenei). The mention of the chador (veil worn by women) also suggest support towards the current government’s ideology in reaction to the Mahsa Amini protests. The tweets from other groups also articulate some opposition towards government repression.

4. Experiments

We test several LLMs and fine-tuned LMs to detect extreme political views of Persian Twitter users based on their biographical information and tweet content. This section is divided into three parts: 1. evaluation of GPT-3.5; 2. comparison with other LLMs; and 3. comparison with fine-tuned classifiers.

4.1. Evaluation of GPT-3.5

Here, we assess GPT-3.5, one of the most prominent conversational LLMs, which has gained a significant amount of attention since it was released to the public in 2022 (Radford et al., 2021). As a multi-lingual generative model (Brown et al., 2020), GPT-3.5 includes the Persian (Farsi) language, which constitutes 0.00856% of its training set, corresponding to a corpus totalling 16,731,301 words.[1] This model has been trained on large datasets of conversation data, including social media posts, customer support interactions, and chatbot logs (Dwivedi et al., 2023). It also employs Reinforcement Learning from Human Feedback (RLHF). With RLHF, the feedback obtained from human evaluators is used to train the model further to maximize the reward received when the generated text aligns with human expectations. (Lambert et al., 2022). For all our experiments, we use OpenAI’s API with the `GPT-3.5-turbo` (September 15^{th}) model with its temperature set at 0 to ensure reproducible results.

The prompts provided to the LLMs consist of two essential components: the task specification and the associated input data. A question is formulated by defining the specific task that the LLM is expected to perform and by providing the input data relevant for that task. Subsequently, the question and the input are concatenated into a single

[1] `https://rb.gy/y2w1t`

Table 1: English Translations of Example (Paraphrased) Tweets Across Groups

Group	Translation
Monarchy Supporters	A great slogan that emerged in the heart of Iran, Zahedan: #JavidShah #Mahsa_Amini During the Iranian freedom-loving march in Berlin, the lion and sun flag was raised. -Saturday, October 23, 2022 #Mahsa_Amini #IranRevolution
Government Supporters	The chador (veil) you have put on is around the enemy's neck. So hold your chador tighter! #Labbaik_Ya_Khamenei #End_of_Appeasement A student who was martyred due to knife attacks by street thugs and hooligans. #Labbaik_Ya_Khamenei #End_Immorality
Other Groups	We will not back down because of the blood you shed and the children you imprisoned. #Mahsa_Amini #Mehrsa_Mousavi We are the voice of years of coercion, suppression, and censorship. #Nation-wide_Strikes #OpIran #Mahsa_Amini

prompt, which is then presented to GPT-3.5 for processing. We explain the task design in detail below.

4.1.1. Designing task description

Language: We initially crafted task descriptions for GPT-3.5 in both English and Persian. These task descriptions were developed by the authors who are native speakers or fluent in both English and Persian. To expand our investigation further, we explored the results provided by the model when instructing GPT-3.5 to translate the Persian task description into English and then using that translated task description to execute the task. In this experiment, our goal is to compare prompts written in Persian with those translated into English or originally written in English.

Level of Details: In another set of experiments, we explored different levels of detail in the questions presented to the model. We came up with three settings: a generic question, a more detailed one, and one with an extensive explanation. The first prompt, 'Generic', only provides the labels "Pro-Government", "Pro-Monarchy" and "Others". The second prompt, 'Detailed', provides some context and defines what the two main classes represent, *"Group 1 supports the Monarchists and demands the restoration of the Pahlavi dynasty. Group 2 stands behind the current Islamic Republic and adheres to strict Islamic laws. Group 3 encompasses all other political stances not falling into these two categories."* The third prompt, 'Extensive', complements this with additional information on the "Others" group by adding, *"Group 3 encompasses all other political stances not falling into these two categories and includes secularists and reformists, women's rights activists and Kurdish activists."* The full list of prompts is provided in Table 2.

4.1.2. Input Design

Given that our dataset includes hyper-partisans who express their political leaning in their bio descriptions, we began the experiments by using the user's bio as input. We then revised our input to provide the model with a more comprehensive user context by adding the user's tweets as input. To select the tweets to be added to input, we experimented with different methods namely: 1. 'latest', which includes most recent tweets; 2. 'hashtag', which includes tweets with popular hashtag; and 3. 'retweet', which includes most popular tweets.

To select tweets related to popular hashtags, we arranged a list of hashtags used by the user throughout the dataset's time period, ranking them based on frequency of usage. We then chose one tweet associated with each hashtag, beginning with the most frequently used ones, depending on the desired number of input tweets. In cases where a user had few hashtags, we re-generated the list to meet the desired input quantity. If a user did not have enough popular hashtag tweets, we randomly selected additional tweets until we reached the required number.

To select the users' most popular tweets, we ranked each user's tweets by the number of retweets they received up to that point and selected input from the top of this list. It is important to note that our dataset was collected in real-time, so at the time we collected a particular tweet, it had no retweets. But because of the collective nature of this movement on Twitter, most of the tweets were already retweets of other users' tweets, and the Twitter API returned us the number of retweets the original tweet had up to that point in time. Therefore, in our experiments, by the *"number of retweets"*, we mean the number of retweets the original tweets had received. Finally, we combined the bio description with tweets selected with our best tweet sampling strategy and

Table 2: List of English Prompts for Hyper-partisan Ideology Detection Task

Input Type	**Detail Level**	**Task Description**
Bio	Detailed	*We are interested in studying political groups in Iran based on Farsi Twitter. Your task is to analyze the bio description of a Twitter user which is translated to English and predict one of the following groups they are most likely to belong to. Group 1 supports the Monarchists and demands the restoration of the Pahlavi dynasty. Group 2 stands behind the current Islamic Republic and adheres to strict Islamic laws. Group 3 encompasses all other political stances not falling into these two categories. Respond with '1', '2', or '3' with no other text or explanation. \n Bio description: {input_text}*
Bio	Generic	*We are interested in studying political groups on Farsi Twitter. Your task is to analyze the bio description of a Twitter user and predict one of the following groups they are most likely to belong to. Group 1 supports the idea of monarchy. Group 2 stands behind the Islamic Republic. Group 3 encompasses all other political stances not falling into the first two categories. Respond with '1', '2', or '3' with no other text or explanation. \n Bio description: {input_text}*
Bio	Extensive	*We are interested in studying political groups in Iran based on Farsi Twitter. Your task is to analyze the bio description of a Twitter user and predict one of the following groups they are most likely to belong to. Group 1 supports the Monarchists and demands the restoration of the Pahlavi dynasty. Group 2 stands behind the current Islamic Republic and adheres to strict Islamic laws. Group 3 encompasses all other political stances not falling into these two categories and includes secularists and reformists, women's rights activists, and Kurdish activists. Respond with '1', '2', or '3' with no other text or explanation.\n Bio description: {input_text}*
Tweets	Detailed	*We are interested in studying political groups in Iran based on Farsi Twitter. Your task is to analyze the tweets of a Twitter user and predict one of the following groups they are most likely to belong to. Group 1 supports the Monarchists and demands the restoration of the Pahlavi dynasty. Group 2 stands behind the current Islamic Republic and adheres to strict Islamic laws. Group 3 encompasses all other political stances not falling into these two categories. Respond with '1', '2', or '3' with no other text or explanation. \n Tweets: {input_text}*
Bio and Tweets	Detailed	*We are interested in studying political groups in Iran based on Farsi Twitter. Your task is to analyze the bio description and tweets of a Twitter user and predict one of the following groups they are most likely to belong to. Group 1 supports the Monarchists and demands the restoration of the Pahlavi dynasty. Group 2 stands behind the current Islamic Republic and adheres to strict Islamic laws. Group 3 encompasses all other political stances not falling into these two categories. Respond with '1', '2', or '3' with no other text or explanation.\n Bio:{bio} \n Tweets:{tweets}*

gave it to GPT-3.5 as the input of the prompt.

4.2. Evaluation of Open-source models

We compared GPT-3.5 with two open-source LLMs for this task. These models are Llama 2 (70B Chat)[2] and WizardLM (70B)[3]. Llama 2 (Touvron et al., 2023b) is a collection of pre-trained generative text models developed by Meta. The scale of this model is from 7 billion to 70 billion parameters. Meta claims instruction-tuned Llama 2 Chat series is optimized for multi-round dialogue and outperforms open-source chat optimized models on most benchmarks. Meanwhile, building upon the original LLaMA (Touvron et al., 2023a) framework, WizardLM (Xu et al., 2023) elevates LLM with additional functionalities. This model is also fined-tuned for chat using AI-evolved instructions.

The first challenge we encountered with open-source models was defining their specific tasks, which required us to provide more detailed instructions on the desired format of the output. Indeed, our evaluation method was required to receive a numerical label for each user from the language model. However, these open-source models did not follow through and responded with sentences instead of numbers. We attempted to use regu-

[2] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

[3] https://huggingface.co/WizardLM/WizardLM-70B-V1.0

lar expressions to extract the labels from the text. However, this approach did not work because the sentences produced by the models were often too complex. We opted to use an alternative approach for fixing this issue that required passing the models' responses through an LLM once more to code the intended label suggested by the text. This step to generate a numerical label was done using GPT-3.5, which turned out to be remarkably effective.

Two experiments were conducted on these LLMs: the first one involved using the bio as input for the detailed task description; the second one was a detailed task description with the translation of bios to English as input. We translated the bios using GPT-3.5. For the translation task, we use the prompt "*Here is the bio of a user's Twitter account. Translate it into English. Please respond with only the translation and no further explanation. \n Twitter bio: {input_text}*"

Both of the models were run with the temperature set to 0 and with vLLM (Kwon et al., 2023), a framework which uses the PagedAttention to optimize the utilization of GPU memory and improve performance.

4.3. Evaluation of fine-tuned classifiers

Another method that we believe is useful in ideology prediction with textual data implies fine-tuning the classifiers. In this research, we employed two classifiers from the BERT family: ParsBERT for the Persian language and RoBERTa for the English language. We also ran preliminary experiments on multilingual BERT (m-BERT), but the results were much worse than those of RoBERTa and ParsBERT, so we did not continue with this model.

The two input designs that showed the most promising results on GPT-3.5 were given to classifiers separately. These input designs include: 1. bio description of the users; and 2. 20 tweets selected based on the most used hashtags of the users. We then used the data translated by GPT-3.5 to work with RoBERTa and fine-tuned the language models with the following hyper-parameters: ParsBERT with *batch size*=16, *learning rate*=0.0001, *warm-up steps*=0, *weight decay*=0.205, *epochs*=3, and RoBERTa with *batch size*=16, *learning rate*=0.00008, *warm-up steps*=100, *weight decay*=0.251, *epochs*=4.

5. Results and Discussions

Our initial experiments aimed to assess the performance of GPT-3.5 with various approaches. In our first experiments, we observed that GPT-3.5 failed in this task when prompted with a few-shot strategy. This explains why we decided to adopt a zero-shot prompting strategy for the rest of the paper.

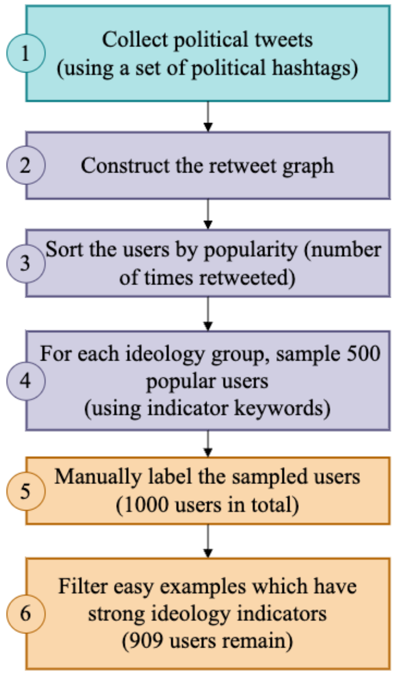
1
Collect political tweets (using a set of political hashtags)
2
Construct the retweet graph
3
Sort the users by popularity (number of times retweeted)
4
For each ideology group, sample 500 popular users (using indicator keywords)
5
Manually label the sampled users (1000 users in total)
6
Filter easy examples which have strong ideology indicators (909 users remain)

Figure 3: Dataset construction process

Prompt	F1	Accuracy
English	**0.72**	0.70
Persian	0.67	0.67
Translated (Fa to En)	0.70	0.69

Table 3: The results of GPT-3.5 in different task description languages.

First, we prompted GPT-3.5 with the 'Generic' task description in Persian, English, and GPT-3.5's translation of the Persian prompt to English. The inputs in these experiments are Persian bios. Table 3 shows the results for this set of analyses. We can see that the model performs better when prompted in the English language compared to the Persian prompt or the translated prompt. This last result explains why we continue the experiments with English prompts.

We then experimented with different level of details in the prompts. As shown in Table 2, the 'Detailed' prompt and 'Generic' prompt differ in the context provided to the model and the explanation of the first and second groups. Results are presented in Table 4. The two prompts display similar f-scores for the bio input, but the generic prompt outperforms the bio and tweets combination. We can see that the gap between the 'Detailed' prompt and the 'Extensive' prompt is more significant on the bio and tweets input. For the remainder of the analyses, unless stated otherwise, all experiments in this study use the 'Detailed' task description.

We subsequently experimented with different input design strategies the results of which are shown in Table 5. We observe that the best method for

	Bio		Bio and Tweets	
Prompt	F1	Accuracy	F1	Accuray
Generic	0.72	0.71	0.56	0.59
Detailed	**0.72**	0.70	**0.67**	0.68
Extensive	0.69	0.68	0.60	0.63

Table 4: The effect of the level of details provided in the prompt.

Bio	Tweets	Count	F1	Accuracy
✓	✗	-	**0.72**	0.70
✗	latest	5	0.44	0.51
		10	0.46	0.52
		20	0.45	0.52
✗	hashtags	5	0.52	0.57
		10	0.51	0.56
		20	0.52	0.57
✓	hashtags	5	0.62	0.63
		10	0.66	0.67
		20	**0.67**	0.68
✗	retweet	5	0.51	0.54
		10	0.52	0.54
		20	0.51	0.55
✓	retweet	5	0.65	0.66
		10	0.66	0.67
		20	0.64	0.64

Table 5: The effect of input in the GPT-3.5 response.

choosing the most informative tweets is associated with choosing the most used hashtags of the user. Considering that we started by scraping tweets with related hashtags for data collection, this strategy could capture the context more effectively than others, such as popular tweets or latest tweets. The best number of tweets included in the input would appear to be between 10 or 20, depending on the method of choosing tweets. However, using only the bio is still more effective.

Table 6 shows the results of our experiments on LLMs, Llama 2 Chat and WizardLM, which demonstrate that GPT-3.5 outperforms these models in our task. The performance of both LLMs is improved when provided with English translations of bios rather than the original Persian versions. WizardLM is outperforming Llama 2 on Persian, but GPT-3.5 still has a significant lead.

Model	Bio	F1	Acc
GPT-3.5	Original	**0.72**	0.70
	Translated	**0.77**	0.76
Llama 2	Original	0.42	0.45
	Translated	0.71	0.71
WizardLM	Original	0.63	0.62
	Translated	0.71	0.70

Table 6: Open source LLMs versus GPT-3.5.

bio		
Model	F1	Accuracy
GPT-3.5-English-Prompt	0.72	0.70
Fine-tuned ParsBERT	0.81	0.82
Fine-tuned RoBERTa	**0.86**	0.87
bio + tweets		
Model	F1	Accuracy
GPT-3.5-English-Prompt	0.67	0.68
Fine-tuned ParsBERT	**0.86**	086
Fine-tuned RoBERTa	0.85	0.85
tweets		
Model	F1	Accuracy
GPT-3.5-English-Prompt	0.52	0.57
Fine-tuned ParsBERT	0.80	0.81
Fine-tuned RoBERTa	**0.85**	0.85

Table 7: Comparison of the fine-tuned language models and GPT-3.5 across different inputs.

Table 7 include the comparison between GPT-3.5 and the fine-tuned classifiers. These results indicate that fine-tuned models outperform GPT-3.5 on the user classification task. RoBERTa shows the best performance with an f1 score of 0.86, while provided with the translation of the bio description. ParsBERT shows the same f1 score when provided with bios and 20 tweets that are chosen by the most-used hashtags of the user, which is the best setting of Table 5. All the results reported in Table 7 which include tweets have the same setting; that is, they correspond to the results of fine-tuned classifiers when trained and tested with only 20 tweets selected based on hashtags. Also, all of the reported f1 scores in the tables are weighted f1 scores.

From Table 7, we also see that including tweets in the input boosts ParsBERT's performance but adversely affects RoBERTa's performance. We observe that higher accuracy in tweet classification is achieved when tweets are translated to English, and a RoBERTa model is fine-tuned, rather than when using a fine-tuning ParsBERT with Persian tweets directly. This indicates that using RoBERTa on translated text for identifying the ideology of users results in the best performance, especially when the model is only based on their tweets.

Since RoBERTa demonstrates superior performance in Table 7, we conducted an additional qualitative analysis to better understand its performance compared to GPT-3.5. This involved reviewing instances where each model struggled to identify discrepancies in their predictions. Our analysis indicates that GPT-3.5 lacks context awareness, leading to incorrect predictions, even when familiar symbols or mottos of the Persian political context are involved. Conversely, in our analysis, RoBERTa appears to struggle with detecting sarcasm, a common element in Twitter communications. Finally, we also note that GPT-3.5's refusal to translate offensive language could imply that RoBERTa is working with less information due to translation losses.

Finally, we ran some additional experiments with the `gpt-4-1106-preview` model as well. Its performance on bios with the detailed task description is considerably better than GPT-3.5 with weighted f1 score and an accuracy level of 0.83. The performance is also improved with the bios that were translated to English using GPT-3.5. In this experiment, the f1 score and accuracy of GPT-4 is 0.82. Finally, the performance also improves when we translate the bios to English with GPT-4. This time we find an f1 score and accuracy of 0.84. While these results are better than GPT-3.5, we did not perform the experiments of this study on GPT-4 because of the expensive price of this model which makes it less practical.

6. Conclusions and Future Work

This study explored the application of LLMs for political ideology detection in the context of Persian Twitter users. Our results confirmed that the best approach to classifying ideology on Persian Twitter is to fine-tune a ParseBERT model with a combination of user biographies and tweets with the most popular hashtags. A RoBERTa model fine-tuned with translated biographies results in the same f-score, but the added cost of translation makes this approach less practical. However, there are several important limits to our analysis.

We acknowledge that this task is much more complex than similar works conducted in English. First, unlike in democratic countries, the broad spectrum of political views beyond the hyper-partisan ideologies is not well-defined in Iran. Second, Persian is a low-resource language, and LLMs are expected to perform worse in this language than in English. For these reasons, we limited our study to a computationally simplified task of hyper-partisan ideology detection by defining our ideological groups according to two extreme views: one that supports the Islamic government; and another that calls for a return of the overthrown monarchy. All other remaining opposition ideological groups were categorized in an "Others" for this analysis.

We evaluated `GPT-3.5-turbo` in different settings and showed that even in this simple computational task, while GPT-3.5 offered convincing results, it significantly performed worse than specialized models, such as fine-tuned RoBERTa and ParsBERT. These results highlight the importance of language-specific models for computational tasks that involve contextual nuances in a non-English space. Our results also confirm that investing in benchmark datasets to evaluate LLMs in non-English languages and non-standard tasks is extremely important. These datasets are crucially important for understanding the capabilities of LLMs; they are also necessary to develop specialized models that address the diverse needs of non-English speakers.

In future work, we intend to explore the landscape of ideology groups within the "Others" category. Given the diverse range of perspectives and the presence of numerous subgroups with complex boundaries within this category, we anticipate the need for a combination of unsupervised and supervised methodologies to effectively map and understand these varied ideological views.

Ethics Statement

We used the Twitter Research API to collect tweets for this study. In order to comply with Twitter's policies and to respect the users' privacy, we will not make the labelled dataset publicly available. However, our data collection methodology can be used by other researchers to explore similar tasks and scenarios.

There is a risk of mislabelling when users are labelled for political ideology based on their social media activities. Here, we mitigate this risk by labeling users that belong to two hyper-partisan groups ("Pro-Government" and "Pro-Monarchy") based on the explicit ideology identifiers found in the users' bio descriptions. Users without these identifiers were categorized as "Others". This implies that users labeled as "Pro-Government" and "Pro-Monarchy" self-identify with these classes publicly and actively engage in political discussions. We do not release any personally identifiable information for any of the users we studied.

Political ideology detection can be potentially misused by malicious actors to influence users, interfere in other countries' elections, or spread misinformation on social media. We emphasize that this task should never be employed to enable targeting of specific users. But there is no security by obscurity here. To counter such malicious uses, it is critical to develop strategies that reduce the spread of misinformation and extreme polarization, minimize the impact of bots, and promote safe and healthy online environments—tasks for which understanding ideology is essential (Pelrine et al., 2023; Tucker et al., 2017). Therefore, by minimizing risks through measures discussed above, such as not releasing identifiable data, political ideology detection research is beneficial to society.

Finally, it is important to note that GPT-3.5 is a closed system with unknown training data and strategies and frequent updates. Because of these factors, it is difficult to fully analyze and contextualize our results. Furthermore, these results may not remain valid for future versions of the model. However, this study is an important initial effort to explore the capabilities and limitations of general-purpose generative systems compared to fine-tuned supervised models for low-resource non-English languages, specifically the Persian language.

7. Bibliographical References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Pooya Azadi and Mohsen B. Mesgaran. 2021. The clash of ideologies on persian twitter. Working Paper 10, Stanford Iran 2040 Project, Stanford University.

Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. 2023. The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, 15(8):260.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Wei Chen, Xiao Zhang, Tengjiao Wang, Bishan Yang, and Yi Li. 2017. Opinion-aware knowledge graph for political ideology detection. In *International Joint Conference on Artificial Intelligence*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. “so what if chatgpt wrote it?” multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642.

ERFI. 2023. The political stance in Iran. `https://erf.institute/stance`.

Mehrdad Farahani. 2020. Albert-persian: A lite bert for self-supervised learning of language representations for the persian language. `https://github.com/m3hrdadfi/albert-persian`.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53:3831–3847.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, Berlin, Germany. Association for Computational Linguistics.

Rouzbeh Ghasemi, Seyed Arad Ashrafi Asli, and Saeedeh Momtazi. 2022. Deep persian sentiment analysis: Cross-lingual training for low-resource languages. *Journal of Information Science*, 48(4):449–462.

Yupeng Gu, Ting Chen, Yizhou Sun, and Bingyu Wang. 2016. Ideology Detection for Twitter Users with Heterogeneous Types of Links. *arXiv e-prints*, page arXiv:1612.08207.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ali Honari and Donya Alinejad. 2022. Online Performance of Civic Participation: What Bot-like Activity in the Persian Language Twittersphere Reveals About Political Manipulation Mechanisms. *Television & New Media*, 23(8):917–938. Publisher: SAGE Publications.

Julie Jiang, Xiang Ren, and Emilio Ferrara. 2022. Retweet-bert: Political leaning detection using language features and information diffusion on social networks. *ArXiv*, abs/2207.08349.

Hossein Kermani. 2023. #MahsaAmini: Iranian Twitter Activism in Times of Computational Propaganda. *Social Movement Studies*, 0(0):1–11. Publisher: Routledge _eprint: https://doi.org/10.1080/14742837.2023.2180354.

Hossein Kermani and Niloofar Hooman. 2022. Hashtag feminism in a blocked context: The mechanisms of unfolding and disrupting #rape on persian twitter. *New Media & Society*, 0(0):14614448221128827.

Hossein Kermani and Amirali Tafreshi. 2023. Walking with bourdieu into twitter communities: an analysis of networked publics struggling on power in iranian twittersphere. *Information, Communication & Society*, 26(8):1653–1674.

Afshin Khashei. 2021. A not-so-dangerous ai in the persian language.

Adel Khorramrouz, Sujan Dutta, and Ashiqur R. KhudaBukhsh. 2023. For Women, Life, Freedom: A Participatory AI-Based Social Web Analysis of a Watershed Moment in Iran's Gender Struggles. ArXiv:2307.03764 [cs].

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. 2022. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*. Https://huggingface.co/blog/rlhf.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ammar Maleki. 2022. IRANIANS' ATTITUDES TOWARD POLITICAL SYSTEMS: A 2022 SURVEY REPORT.

Kellin Pelrine, Anne Imouza, Zachary Yang, Jacob-Junqi Tian, Sacha Lévy, Gabrielle Desrosiers-Brisebois, Aarash Feizi, Cécile Amadoro, André Blais, Jean-François Godbout, and Reihaneh Rabbany. 2023. Party prediction for twitter.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):281–288. Number: 1.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Ilya Sutskever, Rewon Child, Gretchen Krueger, and Jong Wook Kim. 2021. Chat with gpt: Improving language generation and task-oriented dialogue. `https://openai.com/blog/chatgpt-plus`.

Miguel Ángel Rodríguez-García, Soto Montalvo Herranz, and Raquel Martínez Unanue. 2022. Urjc-team at politices 2022: Political ideology prediction using linear classifiers.

Nick Rogers and Jason Jones. 2021. Using twitter bios to measure changes in self-identity: Are americans defining themselves more politically over time. *Journal of Social Computing*, 2.

Alireza Salemi, Emad Kebriaei, Ghazal Neisi Minaei, and Azadeh Shakery. 2021. Arman: Pretraining with semantically selecting and reordering of sentences for persian abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9391–9407.

Mehrnoush Shamsfard. 2019. Challenges and opportunities in processing low resource languages: A study on persian. In *International Conference Language Technologies for All (LT4All)*.

Heydar Soudani, Mohammad Hassan Mojab, and Hamid Beigy. 2022a. Persian natural language inference: A meta-learning approach. *arXiv preprint arXiv:2205.08755*.

Heydar Soudani, Mohammad Hassan Mojab, and Hamid Beigy. 2022b. Persian natural language inference: A meta-learning approach. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4306–4319, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Petter Tornberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Verity Trott. 2021. Networked feminism: counterpublics and the intersectional issues of #metoo. *Feminist Media Studies*, 21(7):1125–1142.

Joshua A Tucker, Yannis Theocharis, Margaret E Roberts, and Pablo Barberá. 2017. From liberation to turmoil: Social media and democracy. *Journal of democracy*, 28(4):46–59.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. WizardLM: Empowering Large Language Models to Follow Complex Instructions. ArXiv:2304.12244 [cs].

Hao Yu, Zachary Yang, Kellin Pelrine, Jean Francois Godbout, and Reihaneh Rabbany. 2023. Open, closed, or small language models for text classification?

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.