A Few Hypocrites: Few-Shot Learning and Subtype Definitions for Detecting Hypocrisy Accusations in Online Climate Change Debates

Paulina Garcia Corral[1], Avishai Green[2], Hendrik Meyer[3*]
Anke Stoll[4], Xiaoyue Yan[5], Myrthe Reuver[6]
[1] Data Science Lab, Hertie School [2] Hebrew University of Jerusalem [3] University of Hamburg
[4] Ilmenau University of Technology [5] University of Zurich [6] Vrije Universiteit Amsterdam
[1] corral@hertie-school.org, [2] avishai.green@mail.huji.ac.il
[3]hendrik.meyer-1@uni-hamburg.de
* The first three authors contributed equally, and are shared first authors.

Abstract

The climate crisis is a salient issue in online discussions, and hypocrisy accusations are a central rhetorical element in these debates. However, for large-scale text analysis, *hypocrisy accusation detection* is an understudied tool, most often defined as a smaller subtask of fallacious argument detection. In this paper, we define hypocrisy accusation detection as an independent task in NLP, and identify different relevant subtypes of hypocrisy accusations. Our Climate Hypocrisy Accusation Corpus (CHAC) consists of 420 Reddit climate debate comments, expert-annotated into two different types of hypocrisy accusations: personal versus political hypocrisy. We evaluate few-shot in-context learning with 6 shots and 3 instruction-tuned Large Language Models (LLMs) for detecting hypocrisy accusations in this dataset. Results indicate that the GPT-4o and Llama-3 models in particular show promise in detecting hypocrisy accusations (F1 reaching 0.68, while previous work shows F1 of 0.44). However, context matters for a complex semantic concept such as hypocrisy accusations, and we find models struggle especially at identifying political hypocrisy accusations compared to personal moral hypocrisy. Our study contributes new insights in hypocrisy detection and climate change discourse, and is a stepping stone for large-scale analysis of hypocrisy accusation in online climate debates.

1 Introduction

Perhaps no accusation is more commonly lobbied in political discourse as that of hypocrisy (Collins, 2018; Thompson, 2004). Allegations of hypocrisy, defined by the Oxford English Dictionary as the assumption "of a false appearance of virtue or goodness" (OED, 2024) and understood in practice as an incongruity between behavior and publicly expressed beliefs (Furia, 2009), are so ubiquitous as to lead Hannah Arendt to describe politics itself as a "never ending fight to ferret out hypocrites" (Arendt, 2006). Making such an accusation of one's rival is effective, as hypocrisy is widely and deeply loathed, with research showing that perceived hypocrisy negatively affects voters' opinions of politicians above and beyond other underlying scandals (Bhatti et al., 2013; Laurent et al., 2014; Grover and Hasel, 2015). It may also seem, in a polarized political landscape, like the only rhetorical tool available. When political opponents lack shared standards, moral persuasion becomes near-impossible, and what is left is undermining one's opponent with "the revelation that... [he or she] is not living up to his own professed ideal" (Shklar, 1984) . While discourse on political hypocrisy dates centuries back, it has only intensified online, buoyed by social media's amplification of polarization (Allcott et al., 2020) and valorization of authenticity (Hallinan et al., 2021).

An arena of online discourse in which the hypocrisy charge is especially pertinent is that of debates around climate change, where hypocrisy accusations have been shown to be central to increasingly polarized, cross-ideological online interactions (Brüggemann and Meyer, 2023). An example of such an online hypocrisy accusation is a Reddit commenter describing COP26, an international political climate conference in 2020, as *"the biggest hypocrisy in the world"* [because the politicians are] *"arriving in their private jets."* Hypocrisy accusations have been found to drive polarization in online climate debates (Falkenberg et al., 2022), facilitating not cross-camp deliberation but rather segregation into opposing ideological camps (Meyer et al., 2023) and affective polarization (Tyagi et al., 2020). Being able to analyze (online) debates about climate policy, focusing on hypocrisy accusations can help us better measure the use of hypocrisy in such debates.

Detecting accusations of hypocrisy is understudied in computational text analysis and Natural Language Processing (NLP). Usually, hypocrisy is only

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 45–60

included as part of broader logical fallacy detection tasks, as in Habernal et al. (2018). However, hypocrisy accusations are a distinct phenomenon: Linguistically, they often contain contrastive conjunctions between clauses to highlight inconsistencies. Semantically, their identification relies on context to understand contradictions.

Earlier work (see Section 2) reports good performance on detecting *logical fallacies* in general, but it is noticeable in the literature that hypocrisy accusation are a difficult and sometimes neglected phenomenon in state-of-the-art NLP. Alhindi et al. (2022) report relatively low performance of 0.43 in macro-F1 for detecting *whataboutism*, a hypocrisy-related fallacy – and only 0.21 in macro-F1 for all fallacies in climate debates. Recently, Piskorski et al. (2023) report a performance of RoBERTa-XLM multilingual detection of the *whataboutism* concept with an F1 of 0.06. However, this concept consists of only 0.05% of their fallacy dataset.

Due to a lack of attention to the specific *hypocrisy accusation* construct, there is a scarcity of data annotated for hypocrisy detection as well as little research of nuances in this construct, such as different types of hypocrisy accusations. Yet hypocrisy allegations are varied, and we believe an account of different kinds will afford a better understanding of the online debate and the performance of models detecting such accusations.

Our contributions to the literature are:
(1) We, unlike previous work, analyze **hypocrisy accusation detection as an individual and nuanced task** particularly relevant for analyzing online climate discourse;
(2) We are also, to our knowledge, the first to define **different types of hypocrisy** accusation for computational analysis, where we differentiate personal moral hypocrisy from political hypocrisy;
(3) Additionally, we release a **dataset**: the Climate Hypocrisy Accusation Corpus (CHAC), with 420 comments annotated by social scientists for the different types of hypocrisy accusations;
(4) We **analyze the potential of Large Language Models** as hypocrisy accusation detectors and evaluate LLMs with the various hypocrisy constructions we find in our data. We note where models struggle with the task, as a building block for future hypocrisy accusation analyses.

This paper is as organized as follows: Section 2 describes previous research on measuring social science constructs using LLMs, including hypocrisy detection. Section 3 introduces our dataset: the different types of hypocrisy, our annotation process, and corpus statistics. We describe our experiments to evaluate the capabilities of different instruction-tuned models on this dataset in Section 4, and present results of these experiments in Section 5. We then reflect on our results and the complexity of our task (Section 6), and conclude our work in Section 7.

2 Background

Large Language Models (LLMs) are models trained on predicting sequences of text. These models are trained in human preferences and instructions, and can perform in-context learning: with a natural language prompt or instruction, these models are able to do tasks such as classifying text examples on a new construct (Brown et al., 2020).

2.1 Promise and Limits of LLMs for Social Science Construct Detection

LLMs show great promise for detecting complex social science constructs in text. Recent comparative analyses highlight their exceptional performance and adaptability across numerous NLP tasks, including but not limited to sentiment analysis, offensive language detection, intent recognition, fake news classification, stance detection, and document classification (Fields et al., 2024). Alizadeh et al. (2024) compared the performance of open-source LLMs in text classification tasks typical for political science research, employing both zero-shot and fine-tuned LLMs for tasks including stance, topic, and relevance classification on news articles and tweets. They concluded that fine-tuning enhances the performance of open-source LLMs, making it preferable to few-shot training with a relatively modest quantity of annotated text. Törnberg (2023) explored the utility of GPT models for annotating political Twitter messages and found that ChatGPT-4o achieved higher accuracy, reliability, and either equal or lower bias compared to human classifiers. It also excelled in annotations requiring contextual reasoning and inference of authorial intent.

However, these models show some potential limitations. Plaza-del Arco et al. (2023) have found LLMs using in-context learning to outperform other NLP models in detecting complex social constructs such as sexist comments and misogynist hate, though they do report prompt brittleness – resulting in less stability of construct detection over slightly different prompt formulations. Additionally, humans remain better than LLMs at editing and improving difficult examples for model training, e.g. in sexist language detection (Sen et al., 2023). This indicates that LLMs are promising for the annotation and detection of social science construct in text, but also show some limitations in subtle understanding of the underlying construct.

2.2 Fallacies and Hypocrisy Accusations

Hypocrisy accusations are a subtle construct, but are not often the main topic in research using state-of-the-art NLP. Habernal et al. (2018) identify fallacies as unfair arguments in debates, 'deceptions in disguise' whose conceptualization goes back to Aristotle (Aristotle, 1909). Their work locates hypocrisy accusations as a subtype of *ad hominem* fallacy, and find a convolutional neutral network is able to detect *ad hominem* fallacies with 0.81 accuracy. Sahai et al. (2021) mention hypocrisy accusations as a common fallacy in online debate, but do not include it as one of the eight fallacy types they detect with neural models with 0.76 accuracy.

Hypocrisy has also been part of other types of tasks. Piskorski et al. (2023) introduce a multilingual dataset with an annotated hypocrisy accusation concept as part of a 'persuasion techniques task'. They also introduce an XLM-RoBERTa model as a baseline. One of the topics in their dataset is the climate change debate. Their appendix reports a performance of the *whataboutism* concept of 0.25% precision, with extremely low recall (0.034%) leading to an F1 of 0.06. However, this concept is only 0.05% of the entire dataset.

More recently, fallacy detection has also been explored with LLMs. The Logical Fallacy Understanding Dataset (LFUD) (Li et al., 2024) was created to evaluate LLMs' capability of logical fallacy understanding. The authors show how this dataset can be fine-tuned to obtain significantly enhanced performance on logical reasoning. On a limited set of logical fallacies (*Against the Person, Appeal to Authority, Appeal to Popularity, Appeal to Emotion, Hasty Generalization, Questionable Cause, and Red Herring*), GPT-4o achieves an accuracy of 0.79, and when used in cases that exclude invalid or unidentified instances, an accuracy of 0.90 (Lim and Perrault, 2024). Additionally, Valdovinos (2023) created a real-time fallacy detection for events such as presidential debates online, which integrates audio transcription models with four fallacy classification models. However, there is no category of hypocrisy accusations in these previous LLM works, nor do they focus on climate debates.

Alhindi et al. (2023) apply LLMs for detecting fallacies in online climate debates, using five existing fallacy datasets as well a new dataset constructed specifically for the task. Their experimental set-up consists of fine-tuning different sizes of the T5 LLM model on five fallacy datasets with different fallacies (and different topics of debate, from COVID-19 to the climate) before using in-context learning with prompts for detecting fallacies in the target dataset. Using this training scheme, they are able to detect *whataboutism* in one dataset with 0.44 accuracy. Their climate dataset contains no hypocrisy class, and sees an average performance of 0.21 in macro-F1 over nine other fallacies. They acknowledge that context is essential for understanding both the climate debate and *whataboutism*.

Thus, hypocrisy accusations have so far not received sufficient attention in either dataset creation or model development. Yet hypocrisy accusation detection is a complex task in its own right, with both semantic and logical context needed for success, and complexity added by several social language factors such as irony and sarcasm. Additionally, earlier results lack a careful evaluation of which different types of hypocrisy accusations can be detected. In our dataset and experiments, we intend to fill this gap by presenting a specialized climate hypocrisy accusation dataset, and an analysis of the performance of currently popular LLMs in detecting hypocrisy accusations.

3 Data

We present a dataset based on the English-language *Reddit European Sustainability Initiatives* corpus released by Reuver et al. (2024). This corpus consists of 2,073 sustainability discussions from between 2017 and 2022 on the Reddit.com subcommunities (Proferes et al., 2021) called *europe*, *europeanunion*, and *europes*, with 46,285 comments. Nearly half (922) of these discussions have at least one comment.

We focus on the comments in this dataset, as they constitute active discourse between users on the identified discussion topics, which are relevant for hypocrisy accusations. This means that our unit of analysis is a comment, which can contain a single or multiple sentences.

3.1 Data Sample and Annotation Process

Our sample selection involves two main strategies. We divided the data into two groups: 1) instances were hypocrisy was explicitly mentioned, by using the regex pattern `hypocr*`, and 2) the remaining data. Subsequently, we randomly selected 300 samples from each group, and consolidated them into a single dataset with 600 samples. The sampling strategy was done due to the relative rarity of the explicit hypocrisy mentions.

The six expert annotators were the authors of this study, all experts in political science, environmental communication science, or (computational) linguistics, which allowed for thorough, high-quality annotations. All participated in a test round to test the annotation scheme and make comments and adjustments. After flagged issues were solved, we then proceeded to annotate the final dataset.

Each expert annotated half of the final dataset, which yielded 3 annotations per sample. This sample size also aligns with existing literature, which suggests that a dataset of this magnitude is generally sufficient for few-shot learning tasks, offering a balanced measure of model performance with respect to human annotation capabilities.

3.1.1 Annotation Scheme

We devised a nested annotation scheme to identify instances of hypocrisy allegations within statements (Q1), and when these are detected, the type of accusation (Q2). First, hypocrisy allegations are coded binarily: *Hypocrisy Accusation/ No accusation*. A statement is considered an allegation of hypocrisy when it does at least one of the following:

- Includes a direct hypocrisy accusation, such as calling someone a hypocrite or describing their actions as hypocritical (e.g., "COP26 is the biggest hypocrisy in the world, arriving in their private jets")
- Highlights a clear inconsistency or contradiction between someone's actions and their stated values, usually in a way which is negatively morally coded (e.g., "Leonardo DiCaprio simply doesn't get it, protecting marine animals and flying private jets at the same time")
- Employs a rhetorical device such as questioning or invoking hypothetical scenarios to indirectly accuse someone of hypocrisy (e.g., "Shouldn't you consider your own actions before instructing us on what needs to be done?")

The codebook further specifies that allegations can target individuals, institutions, or collectives. Second-hand accusations (e.g., "Lucy Dracus said that Obama is a hypocrite") do not constitute an accusation. Allegations can also be expressed through phrases or sayings synonymous with hypocrisy, such as "double standard" or "one rule for thee, another for me".

In cases in which a hypocrisy accusation was detected (Q1 answered positively), the annotator proceeded to identify the type of hypocrisy (Q2). The categorization into hypocrisy types draws from Gunster's typology (Gunster et al., 2018) of climate hypocrisy discourses, which lays out a distinction between types focused on individual (individual-lifestyle outrage and personal reflective discourse) versus institutional (institutional cynicism and calls to action) behavior. Operationalizing these distinctions, we lay out the following categorization:

A) **Personal moral hypocrisy (PMH)**: a gap between personal behavior and professed beliefs.

   Example: "You claim to care about climate change, yet you eat beef."

B) **Political hypocrisy (PH)**: a discrepancy between professed beliefs, values, or ideology and policy or political action.

   Example: "You talk about the importance of climate change but oppose nuclear power."

   Note: this category also includes inconsistencies between different policy positions.

C) **Neither**: We apply this when we cannot decide between A and B, when there are reasons to choose both, or when we think that neither A or B fit.

When determining the type of hypocrisy, the primary consideration is the content of the targeted action, statement or position. For instance, consumer choices typically indicate personal moral hypocrisy, while explicitly political action such as voting or protesting indicates political hypocrisy. If the content is unclear, the type of actor being accused can guide the decision: accusations against nations or

Label	N
Personal Moral Hypocrisy	35
Political Hypocrisy	35
Neither	2
No accusation	221

Table 1: Count summary label distributions of the labelled dataset

governments are usually political, whereas accusations against private citizens are typically personal. Accusations against specific politicians can be either, depending on the content of the allegation.

3.2 Climate Hypocrisy Accusations Corpus

Our *Climate Hypocrisy Accusations Corpus* (CHAC) corpus consists of 420 labeled comments. We calculate an inter-annotator agreement score of Fleiss' κ: 0.512, indicating a reasonable level of consensus among the annotators. We use majority voting to assign labels to each comment. There are 293 comments with a majority-class assigned label (the rest did not have a majority label). However, we keep the comments without majority consensus and release it with our dataset, as recent calls for perspectivism (Röttger et al., 2022; Romberg, 2022) have highlighted the importance of looking beyond majority consensus when it comes to complex social and argumentative concepts. The distribution of labels is summarized in Table 1. The source-code of the analysis[1] as well as our corpus[2] is available online, released for non-commercial use only under CC-BY-NC licence.

4 Experimental Approach

Our experiments are a first attempt at using the *Climate Hypocrisy Accusations Corpus* to measure the capabilities of different currently popular and high-performing LLMs in detecting hypocrisy accusations. We use few-shot prompting, also known as in-context learning. (Brown et al., 2020). Previous research has shown that for complex social constructs, few-shot out performs zero-shot (Alhindi et al., 2023).

4.1 Model Selection

Our experiments compare two families of high-performing and currently popular LLMs that have shown promise on complex social tasks. We use two GPT series models (Brown et al., 2020) and one LLama series model (Touvron et al., 2023).

We also purposefully chose one more closed and one more open family of models in terms of development and model access. LLama models (Meta AI) do not require payment, and its development team has openly released most of its code and training procedure, while GPT (OpenAI) does require payment and is less open in its architecture. However, almost no currently released LLM by a large technology company is fully open in its release of code, training data, and analysis (see Liesenfeld et al. (2023) for comparing aspects of 'opennness' when it comes to LLMs).[3]

4.2 Prompt and Shot Selection

We opt for a six-shot learning approach to provide the model with two robust examples per category. This provides sufficient context for each classification type without risk of overfitting. This also allows to have some control over the output format, avoiding complex parsing and streamlining the analysis of the results.

We use an iterative prompt design process using the GPT-4o as our base model. We base our prompt on classification formats present in our literature review on fallacy detection. The examples we choose to include in the prompt are not in the test set and are selected to maximize model learning capabilities. This includes hypocrisy examples with complex constructions, reported speech, and rhetorical questions. We also include reasoning in our prompt, as previous literature found reasoning increases model performance, even if it is not reliably correct (Ye and Durrett, 2022).[4]

5 Results

We report our results on the dataset of 293 majority-annotated instances from our *Climate Hypocrisy Accusation Corpus*. To ensure consistency and accuracy in our analysis, we use a systematic parsing process for the strings generated by each model. We standardize the formatting of the outputs to eliminate any discrepancies in punctuation, capitalization, and spacing, using regex patterns to detect and correct common formatting issues, as well as custom scripts designed to handle unique idiosyncrasies of each model's outputs.

[1] https://github.com/pgarco/few-hypo

[2] In https://huggingface.co/datasets/Myrthe/RedditEuropeanSustainabilityInitiatives

[3] Full model description in Appendix A and A.3.

[4] Full prompt is in Appendix A.2

	Acc	Prec	Recall	F1
LLama-3	0.75	0.72	0.71	0.67
GPT-3.5	0.83	0.55	0.49	0.51
GPT-4o	0.75	0.74	0.72	0.68

Table 2: Classification results on the 293 examples labelled for hypocrisy. Results of LLama-3, GPT-3.5, and GPT-4o in accuracy, precision, and recall.

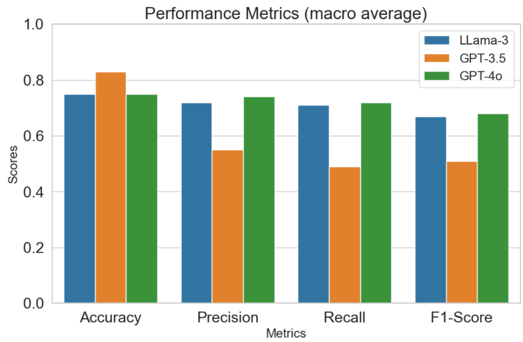

Figure 1: Bar graph comparing result metrics of LLM performance, from left to right we see LLama-3 (blue), GPT-3.5 (orange), and GPT-4o (green), grouped by accuracy (first group), precision (second group), recall (third group), and F1-score (last group).

5.1 Overall Results

The classification results can be seen in Table 2 and Figure 1. LLama-3 and GPT-4o both perform relatively well, significantly outperforming GPT-3.5 overall. In terms of accuracy, all models make predictions that are correct at least 75% of the time, with GPT-3.5 actually leading the way (83%). However, this high accuracy in GPT-3.5 appears to be an artefact of the imbalance between categories. As Table 6 and Figure 2 show, the "No accusation" label is far more prevalent than the other categories, and GPT-3.5 does a better job at predicting it, while GPT-4o and LLama-3 under-predict this class but are better in detecting the two hypocrisy classes.

5.2 Sub-class Prediction

The important difference between the models lies in the prediction of the two hypocrisy classes, reflected in both precision and recall. While GPT-3.5 managed both tasks roughly half of the time, LLama-3 and GPT-4o both succeeded at both tasks above 70% of the time.

Results of the sub-class predictions are visible in Table 3 for GPT-4o , Table 4 for GPT-3.5, and Table 5 for Llama-3. All models perform worse when identifying both Personal Moral Hypocrisy (PMH) (F1 scores between 0.67 and 0.63) and Political Hypocrisy (PH) (F1 scores between 0.46 and 0.54) than in identifying "no accusation" (0.91). All overpredict the different subtype labels, and are worst at identifying accusations not falling under either subtype, though these accusations are very rare. Overall, LLama-3 and GPT-4o respectively have nearly-identical macro-averaged F1 scores of 0.67 and 0.68 over all classes.

We view these scores as a good benchmark for complex hypocrisy accusation detection. The LLama-3 and GPT-4o models show potential for identification of hypocrisy accusations and classification of specific hypocrisy types.

	precision	recall	f1-score	support
No accusation	0.98	0.73	0.84	221
Personal Moral Hypocrisy	0.60	0.74	0.67	35
Political Hypocrisy	0.38	0.91	0.54	35
Neither	1.00	0.50	0.67	2
accuracy			0.75	293
macro avg	0.74	0.72	0.68	293
weighted avg	0.87	0.75	0.78	293

Table 3: Multiclass classification results for GPT-4o on the 293 examples with a majority label for hypocrisy acccusations

	precision	recall	f1-score	support
No accusation	0.87	0.95	0.91	221
Personal Moral Hypocrisy	0.71	0.63	0.67	35
Political Hypocrisy	0.62	0.37	0.46	35
Neither	0.00	0.00	0.00	2
accuracy			0.83	293
macro avg	0.55	0.49	0.51	293
weighted avg	0.82	0.83	0.82	293

Table 4: Multiclass classification results for GPT-3.5-turbo-1025 on the 293 examples with a majority label for hypocrisy acccusations.

	precision	recall	f1-score	support
No accusation	0.97	0.74	0.84	221
Personal Moral Hypocrisy	0.50	0.86	0.63	35
Political Hypocrisy	0.41	0.74	0.53	35
Neither	1.00	0.50	0.67	2
accuracy			0.75	293
macro avg	0.72	0.71	0.67	293
weighted avg	0.85	0.75	0.77	293

Table 5: Multiclass classification results for Llama-3-70b-chat-hf on the 293 examples with a majority label for hypocrisy acccusations.

Label	GPT-3.5	GPT-4o	LLama-3	CHAC
Personal Moral Hypocrisy	31	43	60	35
Political Hypocrisy	21	84	64	35
Neither	2	1	1	2
No accusation	239	165	168	221

Table 6: Distribution of Class Prediction on the 293 examples with a majority label for hypocrisy acccusations.

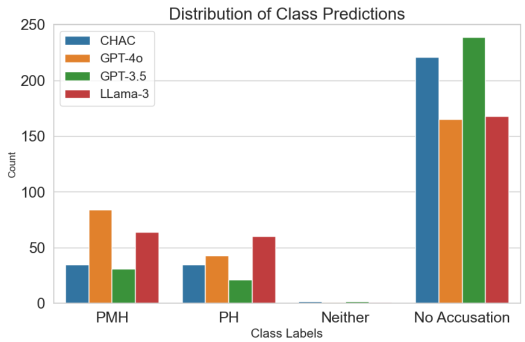

Figure 2: Bar graph comparing prediction and real labels distribution: from left to right we see CHAC dataset (blue), GPT-4o (orange), GPT-3.5 (green) and Llama-3 (red) grouped by class label: PMH (first group), PH (second group), Neither (third group), and No accusation (last group).

5.3 Error Analysis

We perform an error analysis to see whether there are error patterns relating to model type in hypocrisy accusation detection. Overall, we do not find a connection between error type and model family. Each model makes distinct errors.[5]

Broadly, we see three types of error in our results. Firstly, the LLMs predict False Positives, which we refer to as **hypocrisy accusation hallucinations**, often accompanied by a false reasoning in the explanation of the labelling decision generated by the model. A second common error is misclassification of subtype: a correct identification of an accusation, but an incorrect classification of the accusation type. Less common are False Negatives, where a hypocrisy accusation is not found where there is one. Below we discuss each type of error in the different models.

False Positives: Accusation Hallucination While all models sometimes identify hypocrisy accusations where there are none, we find different patterns between the models. GPT-4o and LLama-3 overpredict accusations where there are none, predicting 59 and 58 false positives, respectively. GPT-3.5, meanwhile, hallucinated accusations far less, with only 12 cases of false positives.

Further investigating false positives, we find the cases of hypocrisy accusation hallucination that include the regex pattern `hypocr*` to determine if the presence of the *mention* of hypocrisy could explain the models' over-prediction. Not every mention of the word hypocrisy actually contains a hypocrisy accusation, but it may be a common confusion for models. We find that 37 out of 59 (62%) of GPT-4o's false positives contain the pattern `hypocr*`. For the Llama-3 model, 31 cases (52%), and for the GPT-3.5 errors only 9 cases out of 28 (32%) match the regex pattern. Thus, we observe that GPT-3.5 is better at distinguishing a hypocrisy *accusation* from the *mention* of hypocrisy than other models. See the example below.[6]

> *"What will happen is when it is declared a national emergency, the right will call the left hypocrites for only caring because it was Trump. When in reality, the right's emergency is a baseless claim and climate change is* fucking *real."*
>
> Human label: **Not an accusation**
>
> Llama-3 predicted label: **Political Hypocrisy**
>
> Error: This is a case of the model classifying a second-hand accusation (and a hypothetical one at that) as a hypocrisy allegation. We would not wish to classify this as an allegation: the comment is reporting on this purported accusation in order to refute it.

> *"By that logic basically everyone who wants to stop climate change is a hypocrite."*
>
> Human label: **Not an accusation**
>
> GPT-4o predicted label: **Personal Moral Hypocrisy**
>
> Error: We understand this comment to mention a hypocrisy accusation critically, presumably responding to a hypocrisy allegation by pointing out how its reasoning leads to a conclusion that is *prima facie* absurd. The GPT-4o model seemed to simply take the comment at face value, reasoning that "The commenter is suggesting that anyone who advocates for stopping climate change is a hypocrite". This points to the complexity of understanding nuanced concepts such as irony or sarcasm.

[5]Confusion Matrices for all models are in Appendix B.

[6]Appendix C has examples of the different types of errors.

False Negatives GPT-3.5 misclassifies 28 comments as false negatives – not correctly identifying an accusation. False negatives, then, are the main source of errors for GPT-3.5.

We again look into whether there are any confusions between the *mention* of the word 'hypocrisy' and an *accusation* of hypocrisy. We find that GPT-4o did not label any comment that contains the regex pattern `hypocr*` as *not* being accusations when these were positive cases, while for Llama-3 we find 3 cases of such false negatives.

GPT-3.5 labeled 16 false negative comments containing the pattern. This indicates that GPT-3.5 is not relying on the presence of `hypocr*` related words to label comments, compared to the two other larger models, which is a surprising finding. However, as we see, this did not lead to a better overall performance for GPT-3.5 in detection hypocrisy accusations.

Subtype Misclassification Aside from confusing non-accusations with accusations, the models also show confusion in the subtypes of hypocrisy. We find that these errors are not consistent across models. In the case of *Personal* hypocrisy being labeled as *Political*, the biggest confusion comes from GPT-4o (22%). While Llama-3 has the opposite error: *Political* is labeled as *Personal* (17%). Again, our findings indicate that the smaller model, GPT-3.5, does not confuse types of hypocrisy as much, with only 4 instances of *Political* predicted as *Personal*, and one single case of the opposite. While GPT-3.5 under-predicted the hypocrisy accusations (having more false negatives), it was better at distinguishing between the hypocrisy classes once they were labeled as accusations.

> *"It's James Shaw, the biggest hypocrite out there when it comes to travel. Don't expect a realistic response."*
>
> Human label: **Personal Moral Hypocrisy**
>
> GPT-4o predicted label: **Political Hypocrisy**
>
> Error: The model appears to conclude that James Shaw, the former New Zealand Minister for Climate Change, is a political figure. However, we have noted that accusations against specific politicians can be personal or political depending on their content, and would classify personal travel as a consumer choice and thus an example of personal hypocrisy. However, we recognize this is something of a gray area, and that the model and human coders disagree on it.

We want to find the source of confusion of classes, and investigate if the mention of political figures or events leads to mislabeling. A qualitative analysis reveals that all subtype misclassification cases for GPT-4o include references to political figures or events. In this case, correct annotation of subtype (political hypocrisy or not) requires the model to establish whether the comment has a political context. The errors in LLama-3 and GPT-3.5 of identifying *Political Hypocrisy* as *Personal Moral Hypocrisy* all have references to political figures or events. All four subtype errors by GPT-3.5 are a subset of the error cases made by LLama-3. However, we cannot conclude that references to political figures or events are the source of subclass error, but it is possible that the model is unable to correctly identify mentions as political.

In summary, our error analysis shows some interesting patterns. GPT-3.5, our least successful model overall, is conservative in labelling comments as hypocrisy accusations, but therefore also has less false positives than the other models. We furthermore find this model is less confused by the mention of the word 'hypocrisy'. The other two models, LLama-3 and GPT4o, are better at labelling comments, but also confuse the two subtypes of hypocrisy more. Overall, political hypocrisy seems hard to identify for all models, and the models struggle with identifying a political context from the mention of political actors.

6 Discussion

Our results provide more insight into the complexities of the hypocrisy accusation as a construct. Specifically, we find different model strengths (see Section 6.1) and specifics of different types of hypocrisy accusations (Section 6.2).

6.1 Model Difference

Our results indicate that the most recent models, including GPT-4o, are considerably better than the earlier GPT3.5. This indicates that the current development of instruction-tuned models is one where improvement also means being better at detecting and annotating a complex construct like hypocrisy accusations. Newer models seem to be better at identifying nuanced context. One possible reason for this may be more recently updated training data e.g. new politicians or political events, required to fully understand hypocrisy accusations. However, due to lack of model training transparency, this cannot be verified.

Additionally, the acceptable results of both LLama-3 and GPT-4o are worth considering in light of the former being more open source and freely available, and the latter a more closed model requiring payment. Given their similar performance, the more open model offers clear advantages to researchers in social science. More open models are better for science: these are more reproducible, understandable and allow researchers to not be dependent on paying a third party (Liesenfeld et al., 2023). While none of our tested models are completely open in all aspects (training data, code, and openly accessible), it is useful for social scientists to know that Llama is a LLM that is easily accessible, able to analyze a complex construct such as a climate hypocrisy accusation, and not requiring third-party payment.

6.2 Results in the Context of Other Research

The results also suggest that hypocrisy detection, as most logical fallacies, is a complex task. A pattern-matching approach would consider mentions of 'hypocrite' as positive cases, but – as our experiments illustrate – such accusations also occur indirectly, and their detection is complicated by reported speech, sarcasm, rhetorical questions, and other devices in online debates. We observe that using LLM in-context prompting, detection of hypocrisy accusations achieves decent results. However, compared to other classification tasks for social science constructs (Lim and Perrault, 2024), using LLMs to classify accusations of hypocrisy is below expected model capabilities.

Our results (Macro F1 = 0.68 for GPT-4o) see a performance gap with the fallacy detection on other fallacies reported in earlier work, e.g. F1 = 0.76 for fallacy accusation detection in Sahai et al. (2021), and 0.81 in Habernal et al. (2018) for detecting *ad hominem* attacks. However, these earlier papers did not distinguish hypocrisy accusations from other fallacies, and this narrowing of the concept could lead to more difficulty for our models.

Other literature also reports a more mixed performance of LLMs (especially when compared to fine-tuned Transformer models) for fallacy detection. Ruiz-Dolz and Lawrence (2023) report a F1 score of 0.79 by a fine-tuned Roberta model for two fallacy argument datasets with classes such as *ad hominem attacks* and *appeal to majority*, and in contrast a 0.56 F1 score for GPT-4o on these datasets. This paper also reports a lower performance of GPT-4o on the *ad hominem* fallacy class. Potentially, this could be because of a connection to hypocrisy accusations, which (as we have established) are difficult, and are often forms of *ad hominem* arguments, criticizing the rival personally instead of their positions.

Our results indicate that hypocrisy accusation is an interesting concept that deserves its own task, as well as benchmark datasets outside of more common fallacy datasets. Moreover, these results prove the usefulness of breaking down complex constructs into sub-categories: we found that detecting the hypocrisy/no hypocrisy distinction is relatively easy (e.g., reaching F1 > 0.80 for all models), while the subclass Political Hypocrisy is much harder to detect, showing F1s <= 0.50s for all models. The subclass that may be especially relevant for political analysis, attacks on political actions or views, is not well-detected by LLMs.

7 Conclusions

Hypocrisy accusations are central to increasingly polarized, cross-ideological online interactions. Despite recent research on detecting argument fallacies, hypocrisy accusations remain under-researched and are often a small sub-class in argument fallacy datasets. We define hypocrisy accusation detection as an individual NLP task and create an annotation scheme where we identify sub-classes of hypocrisy accusation: personal moral hypocrisy versus political hypocrisy. We present a dataset, the *Climate Hypocrisy Accusation Corpus* (CHAC), consisting of 420 reddit comments, annotated by six experts. Using our dataset, we compare three different instruction-tuned models (GPT-4o , GPT-3.5, and Llama-3) in a six-shot setting for detecting hypocrisy accusations. The different models have different strengths, but overall perform with a macro F1 class of around 0.80, and show that Llama, as a more open model than the GPT-family and one not requiring payment, can perform on par for hypocrisy accusation detection with the less open GPT model that requires payment. LLMs are capable of detecting accusations with a binary distinction, but we identify room for improvement when it comes to the different accusation types. Models are somewhat worse at detecting political hypocrisy than personal moral hypocrisy, which could have implications for social science research.

Limitations

As with all research, this paper has some limitations. We identify four sources of limitations.

Data Annotation First, the data annotation process could have led to higher inter-annotator agreement score, which, research has shown, is detrimental to achieving high results in computational modelling. However, considering this is a complex theoretical construct, we are satisfied with this first limited result.

Debate Context Second, the validity of the results should be understood as pertaining especially to climate change discourse; As we have described in the paper, hypocrisy relies on the understanding of a contrast between two events, usually a professed belief and an action. These actions and beliefs often need to be understood in context in order to be understood as (allegedly) inconsistent. Hence, we expect identification of hypocrisy accusations to be somewhat dependent on an understanding of both the factual reality of a topic as well as the social context in which it is discussed. This paper analyzes the climate change debate, and the tool's relevance to other fields requires further study and, potentially, training.

Geographical, Linguistic, and Cultural Context Additionally, we acknowledge that our paper is focused on European debates around climate change, in a single high-resource language, English (Bender, 2019). The results we find depend on the data the models were trained on, and as such we expect that non-European debates and debates in low-resource languages will probably produce results that are not as high.

Political Context Lastly, LLMs are not without its issues for social science analysis: these models display political worldviews (Ceron et al., 2024). When analyzing different political contexts (e.g. one more conservative than European climate debates), the results could therefore differ. While adding these results is beyond the scope of the current paper, it is important to keep this in mind for future work.

Ethics Statement

The data used in this project was scraped from Reddit in December 2022 with the PushShift API, before Reddit's PushShift API restrictions were enforced in April 2023, ensuring compliance with the platform's terms of service at the time. We remove any personal identifying information such as usernames from the data. We also ensure the data is released for non-commercial use only. This is also in-line with Reddit users' concern of their data being used for training commercial LLMs or other technology.

Furthermore, some comments reflect personal opinions that are not in-line with the established scientific consensus on climate change. While these opinions are valuable for understanding public sentiment and discourse, we do not endorse any misinformation or scientifically inaccurate statements present in the dataset. Our goal is to analyze these discussions to better understand the dynamics of public discourse on climate change, promoting more effective strategies to engage with the public, address misconceptions, and promote scientifically accurate information.

Acknowledgements

The dataset used in this research was funded by a research voucher grant for the project *Reasons for online (dis)trust in sustainable initiatives* awarded to Myrthe Reuver and Ana Isabel Lopes by the Network Institute at the Vrije Universiteit Amsterdam. Myrthe Reuver was also funded by the Netherlands Organisation for Scientific Research (NWO) through the *Rethinking News Algorithms* project (grant nr: 406.D1.19.073). Paulina Garcia Corral was funded by the DFG (EXC number 2055 – Project number 390715649, SCRIPTS).

A preliminary version of this paper was presented at the hackathon at the International Communication Association (ICA) 2023 conference in Toronto, where this project was initiatied and where all authors participated, and another preliminary version was presented at CompText 2024 in Amsterdam. We would like to thank the ICA hackathon organizers and participants as well as all reviewers, whose comments improved this paper.

We used Github Copilot and ChatGPT to assist in writing our experimental code, and editing code snippets for data processing and data analysis. We used ChatGPT to assist in writing sections of this manuscript. This assistance, in accordance with the ACL Ethics Policy, was solely in dealing with the language of the paper rather than in producing new content or new ideas. All final writing is ultimately done by the authors, who are responsible for it.

References

2024. Oxford english dictionary. Accessed June 19, 2024.

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8172–8187, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2023. Multitask instruction-based prompting for fallacy recognition. *arXiv Preprint arXiv:2301.09992*.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2024. Open-source llms for text annotation: A practical guide for model setting and fine-tuning.

Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The welfare effects of social media. *American Economic Review*, 110(3):629–76.

Hannah Arendt. 2006. *On Revolution*. Penguin.

Aristotle. 1909. *The Rhetoric of Aristotle: A Translation*. Cambridge University Press, Cambridge [Cambridgeshire].

Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14:34.

Yosef Bhatti, Kasper M. Hansen, and Asmus Leth Olsen. 2013. Political hypocrisy: The effect of political scandals on candidate evaluations. *Acta Politica*, 48(4):408–28.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Michael Brüggemann and Hendrik Meyer. 2023. When debates break apart: discursive polarization as a multi-dimensional divergence emerging in and through communication. *Communication Theory*, 33(2-3):132–142.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in llms. *arXiv preprint arXiv:2402.17649*.

Timothy P. Collins. 2018. *Hypocrisy in American Political Attitudes*. Springer International Publishing, Cham.

Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, and Walter Quattrociocchi. 2022. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–21.

John Fields, Kevin Chovanec, and Praveen Madiraju. 2024. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, 12:6518–6531.

Peter A. Furia. 2009. Democratic citizenship and the hypocrisy of leaders. *Polity*, 41(1):113–133.

Steven L. Grover and Marcus C. Hasel. 2015. How leaders recover (or not) from publicized sex scandals. *Journal of Business Ethics*, 129(1):177–94.

Shane Gunster, Darren Fleet, Matthew Paterson, and Paul Saurette. 2018. 'why don't you act like you believe it?': Competing visions of climate hypocrisy. *Frontiers in Communication*, 3:49.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. *arXiv Preprint arXiv:1802.06613*.

Blake Hallinan, Bumsoo Kim, Rebecca Scharlach, and Tommaso Trillò. 2021. Mapping the transnational imaginary of social media genres. *New Media & Society*.

Sean M. Laurent, Brian A. M. Clark, Stephannie Walker, and Kimberly D. Wiseman. 2014. Punishing hypocrisy: The roles of hypocrisy and moral emotions in deciding culpability and punishment of criminal and civil moral transgressors. *Cognition and Emotion*, 28(1):59–83.

Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. 2024. Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*, pages 1–6.

Gionnieve Lim and Simon T. Perrault. 2024. Evaluation of an llm in identifying logical fallacies: A call for rigor when adopting llms in hci research.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Hendrik Meyer, Amelia Katelin Peach, Lars Guenther, Hadas Emma Kedar, and Michael Brüggemann. 2023. Between calls for action and narratives of denial: Climate change attention structures on twitter. *Media and Communication*, 11(1):278–292.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Debora Nozza, Dirk Hovy, et al. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics.

Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2):20563051211019004.

Myrthe Reuver, Alessandra Polimeno, Antske Fokkens, and Ana Isabel Lopes. 2024. Topic-specific social science theory in stance detection: a proposal and interdisciplinary pilot study on sustainability initiatives. In *4th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS), co-located with KONVENS*.

Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*. Association for Computational Linguistics.

Saumya Yashmohini Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. In *ACL-IJCNLP 2021-Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of llm-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504.

Judith N. Shklar. 1984. *Ordinary Vices*. Belknap Press of Harvard University Press.

Dennis F. Thompson, editor. 2004. *Restoring Responsibility: Ethics in Government, Business, and Healthcare*. Cambridge University Press, Cambridge.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *ArXiv*, abs/2304.06588.

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

Aman Tyagi, Joshua Uyheng, and Kathleen M. Carley. 2020. Affective polarization in online climate change discourse on twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 443–447.

Lino Valdovinos. 2023. Real-time fallacy detection. `https://github.com/latent-variable/Real_time_fallacy_detection`.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.

Appendix

A Implementation Details

A.1 Model Details

- GPT-3.5-turbo-0125: This is an OpenAI GPT series (Brown et al., 2020) model, optimized for efficient and cost-effective performance in conversational AI tasks, providing advanced language understanding and generation capabilities. This is a closed-source model.

- GPT-4-turbo: An enhanced version of OpenAI's GPT-4, this model offers improved speed and performance for complex language processing tasks, making it ideal for both conversational agents and other sophisticated AI applications. This is a closed-source model.

- Llama-3-70b-instruct[7]: Developed by Meta AI (Meta, 2024), this is a language model with 70 billion parameters, designed for high-quality conversational AI, capable of understanding and generating human-like text in diverse contexts. This is a non-proprietary model, i.e., it does not require payment.

A.2 Prompt

System:
You are an advanced classification AI. Your task is to labels Reddit comments following the instructions below:
Instructions
Accusations of hypocrisy are when someone accuses someone or something of being a hypocrite because of a moral contradiction between what they profess and their actions or words. Accusations of hypocrisy are sometimes direct such as "You are a hypocrite" and sometimes are understood via context, by highlighting a connoted inconsistency between what people profess or claim their believe and what people actually do. An example would be criticizing a celebrity for using a private jet while publicly claiming the need for climate action.
There are different types of hypocrisy accusations, such as accusing someone of personal moral hypocrisy or accusing someone of political hypocrisy. The primary consideration to decide the accusation type is the content of the targeted action, statement or position. For instance, consumer choices typically indicate personal moral hypocrisy, while explicit political action such as voting or protesting indicates political hypocrisy. If the content is unclear, the type of actor being accused can guide the decision: accusations against nations or governments are usually political, whereas accusations against private citizens are typically personal.
Classification steps
1. You will analyze a Reddit Comment and evaluate if there is an accusation of hypocrisy.
2. You will make sure to consider contextual factors, implications, and rhetorical devices to make sure things such as irony or double meaning are taken into account.
3. You will assign a label to the comment: "Not an accusation" or "Hypocrisy accusation".
4. If the label is "Hypocrisy accusation", then you will analyze what type of accusation it is:
1) "Personal moral hypocrisy": a gap between personal behavior and professed beliefs
2) "Political hypocrisy": a gap between professed beliefs, values, or ideology and policy or political action.
3) "Neither"
5. Then you will assign a second label to the hypocrisy accusations: "Personal moral hypocrisy", "Political hypocrisy" or "Neither".
Choose the most appropriate label, and remember to consider all the rhetorical devices.
User:
Label the next comment:
Not being vegan seems kinda hypocritical to me, if you're serious about climate change and have looked into it then you'd know that it's honestly the easiest way we can make a big impact as consumers. Vegan activists aren't hurting anyone or being violent. Meanwhile, farming is directly causing pain, suffering, and death to animals for profit. That's the real difference.
Assistant:
Label: Hypocrisy accusation - Personal moral hypocrisy
Reasoning: The commenter is directly stating that it is hypocritical to care about climate change and not be vegan. This identifies a perceived gap between the professed belief in fighting climate change and the personal behavior of not adopting a vegan lifestyle. The hypocrisy accusation here is based on an individual's personal moral stance

[7] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

(caring about climate change) not aligning with their personal behavior (not being vegan).
User:
Label the next comment:
For real?? Are you saying their hypocrites because they protest carbon mining? I've seen the mines firsthand they're wrecking nature, displacing people, and demolishing villages just for that dirty, low-energy rock. Get it right, it's the most disgusting and destructive form of electricity generation ever.
Assistant:
Label: Not an accusation
Reasoning: The invocation of hypocrisy in the second sentence ("Hypocrites" because. . . ?") is not itself an accusation, but a rhetorical questioning of such an allegation, one which the commenter subsequently answers negatively ("Smh", meaning "shaking my head", i.e. expressing incredulity at the accusation). The commenter goes on to express frustration and criticism towards the use of lignite for electricity generation, highlighting its environmental impact and the availability of other energy sources. Bottom line: the commenter is defending someone from a hypocrisy accusation.
User:
Label the next comment:
Climate activists think they're on a moral high horse, but they're just making life tougher for everyone else. We've got jobs to get to, and they claim they're saving the future but end up screwing up our daily commutes. No one feels sorry for them—that's not how you protest effectively.
Assistant:
Label: Hypocrisy accusation - Political hypocrisy
Reasoning: The commenter is accusing climate activists of claiming to care about the future of humanity while making daily life harder for others, specifically through disrupting commutes. This highlights a perceived inconsistency between their professed goal (caring for humanity's future) and their actions (disrupting daily life). The hypocrisy accusation is focused on the contradiction between the activists' political goal (caring about the future of humanity) not aligning with the direct results results their behavior (disrupting commutes and making life harder for others).
User:
Label the next comment:
Ugh, this "Green" parties are pathetic. They choose Brown coal over nuclear energy, but won't shut up about air pollution. Get a grip.
Assistant:
Label: Hypocrisy accusation - Political hypocrisy
Reasoning: The commenter is highlighting a (purported) inconsistency between what the Greens profess to care about (c̈rying around about air pollution)̈ and what they actually have done (p̈referred Brown coal over nuclear energy)̈. In addition, this alleged contrast carries a clear negative connotation (c̈rying around)̈. The inconsistency being alleged is between professed values (caring about air pollution) and political action (supporting coal over nuclear energy).
User:
Label the next comment:
I read only that he was accused of being not only a war criminal, but also a hypocrite.
Assistant:
Label: Not an accusation
Reasoning: While the comment does report on an explicit hypocrisy accusation, this is an instance of reported speech. We count only hypocrisy allegations made or endorsed by the commenter, not descriptions of accusations made by someone else.
User:
Label the next comment:
I agree that, if it's there, people need to use trains instead of flying.
However, what does it matter if it's hypocritical? Just because *they* are hypocrites and chose to fly, doesn't mean what they say is untrue. The world is complicated, grow up.
Assistant:
Label: Hypocrisy accusation - Personal moral hypocrisy
This comment it likely responding to a hypocrisy accusation. Even though the commenter seeks to defend the people in question from the criticism inherent in such an accusation, they do so by conceding that the hypocrisy allegation is likely true (j̈ust because they are hypocrites)̈ - even if they contend this is not in fact so bad. While hypocrisy allegations in general carry a negative connotation, this is an unusual case in which such an allegation is explicitly made (or conceded) while attempting to divorce it from the (usual expected) negative charge. The inconsistency being alleged in this allegation stems from personal lifestyle choices - choice of commute, i.e. not taking a train.
User:
Label the next comment: `REDDIT COMMENT`

A.3 Training strategy, model parameters, and other resources

For data processing we used Pandas. To test our prompt and run inference we used the OpenAI API for the GPT models, OpenAI provides their own processing equipment and is not specified. For the Meta AI model, we used TogetherAI API for inference. We used 1 Nvidia A100 GPU for our experiment. We used all default model parameters and set the generation parameters to restrict creativity (temperature = 0, max_tokens=200, top_p=1, frequency_penalty=0, presence_penalty=0). To ensure reproducibility, each model's output was subjected to identical preprocessing and post-processing steps. To analyze classification results, we used the "classification_report" function from Scikit-learn.

B Error Analysis Figures

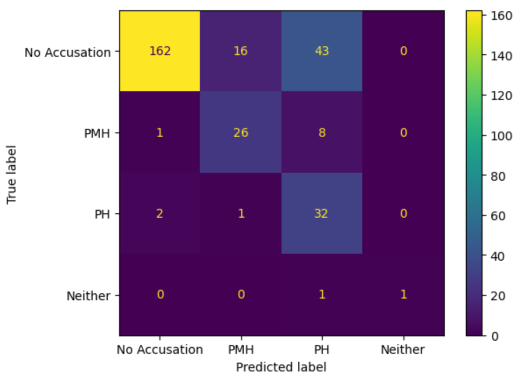

Figure 3: Confusion Matrix for predictions of GPT-4o .

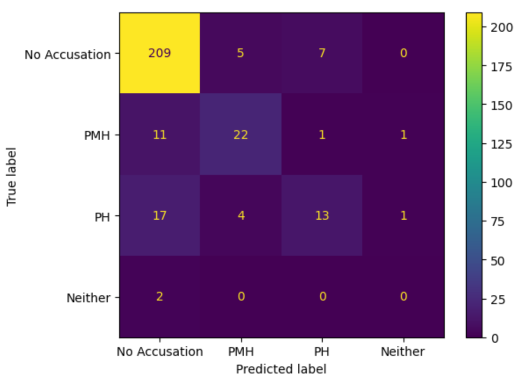

Figure 4: Confusion Matrix for predictions of GPT-3.5.

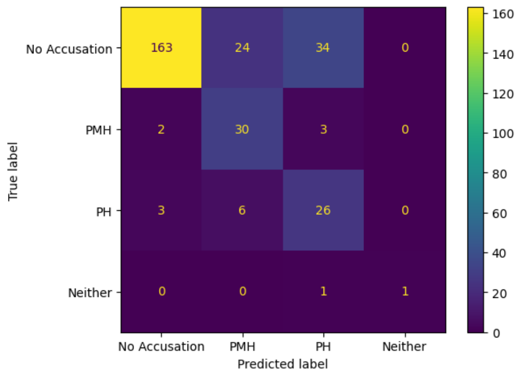

Figure 5: Confusion Matrix for predictions of Llama-3 70B.

C Examples Error Analysis

Examples of Misclassifications

1. *"It's James Shaw, the biggest hypocrite out there when it comes to travel. Don't expect a realistic response."*

   Human label: **Personal Moral Hypocrisy**

   GPT-4o predicted label: **Political Hypocrisy**

   The model appears to based its reasoning on the fact that James Shaw, the former New Zealand Minister for Climate Change, is a political figure. However, we have noted that accusations against specific politicians can be personal or political depending on their content, and would classify personal travel as a consumer choice and thus an example of personal hypocrisy. However, we recognize this is something of a gray area and the model and human coders disagree on it.

2. *"Conservatives are bollocks but let me tell you how good their policies actually are! Hypocrite much?"*

   Human label: **Political Hypocrisy**

   Llama-3 predicted label: **Personal Moral Hypocrisy**

   The comment points to a clear inconsistency between one's political belief (conservatives are bad) and action (praising their policies). However, the GPT-4o model appears to interpret on the fact that the accusation appears directly at another speaker as evidence it is a case personal moral hypocrisy.

Examples of False Positives

1. *“What will happen is when it is declared a national emergency, the right will call the left hypocrites for only caring because it was Trump. When in reality, the right’s emergency is a baseless claim and climate change is* fucking *real.”*

   Human label: **Not an accusation**

   Llama-3 predicted label: **Political Hypocrisy**

   This is a case of the model classifying a second-hand accusation (and a hypothetical one at that) as a hypocrisy allegation. We would not wish to classify this as an allegation, especially as the comment is reporting on this purported accusation in order to refute it.

2. *“By that logic basically everyone who wants to stop climate change is a hypocrite.”*

   Human label: **Not an accusation**

   GPT-4o predicted label: **Personal Moral Hypocrisy**

   While this comment makes an explicit suggestion of hypocrisy, we understand it to be doing so critically, presumably responding to a hypocrisy allegation by pointing out how its reasoning leads to a conclusion that is prima facie absurd. Two of the models concurred, but the GPT-4o model seemed to simply take the comment at face value, reasoning that "The commenter is suggesting that anyone who advocates for stopping climate change is a hypocrite". This points to the complexity of understanding nuanced concepts such as irony or sarcasm.