Influence Reasoning Capabilities of Large Language Models in Social Environments

Luke Gassmann[1], Jimmy Campbell[2], Matthew Edwards[1]
[1]University of Bristol, Bristol, United Kingdom, BS8 1UB
[2]University of Portsmouth, Portsmouth, United Kingdom, PO1 2UP
luke.gassmann@bristol.ac.uk, jimmy.campbell@port.ac.uk, matthew.john.edwards@bristol.ac.uk

Abstract

We ask whether state-of-the-art large language models can provide a viable alternative to human annotators for detecting and explaining behavioural influence online. Working with a large corpus of online interactions retrieved from the social media platform Mastodon, we cross-examine a dataset containing 11,000 LLM influence labels and explanations across nine state-of-the-art large language models from 312 scenarios. We use a range of resolution categories and four stages of shot prompting to further measure the importance of context to language model performance. We also consider the impact of model architecture, and how social media content and features from the explanation impact model labelling accuracy. Our experiment shows that whilst most large language models struggle to identify the correct framing of influence from an interaction, at lower label resolutions, models like Flan and GPT-4 Turbo perform with an accuracy of 70%-80%, demonstrating encouraging potential for future social influence identification and explanation, and contributing to our understanding of the general social reasoning capabilities of large language models.

Introduction

The creation of human language provided our species with a tool to convey and transfer information at a complexity otherwise unseen (Manning 2022). A key application of language is for individuals to exert influence over the beliefs and behaviours of others. With the growth of social media, the average person's capacity for exerting influence has grown from being able to reach a handful of targets to potentially millions. Whilst the discussion of commonsense/-social reasoning in machines has predated Large Language Models (LLMs) by 50 years (Bar-Hillel 1960), the advent of the transformer architecture and its wide array of capabilities has accelerated this topic into scientific focus (Shen and Kejriwal 2023; Gassmann, McConville, and Edwards 2023). In particular, if machines could replicate the human *world model* (Hu and Shu 2023) and demonstrate compelling *theory of mind* (ToM), the implications for machine-human interaction could be significant (Gandhi et al. 2023; Sap et al. 2019a).

One task for which LLMs capable of performing social reasoning would be particularly desirable arises from scientific research studying human interactions. Human annotation is often considered the *gold standard* or *ground truth* for labelling such data (Geiger et al. 2021), however, in many cases, human labels are imperfect. Biases, affective variability and simple boredom can significantly affect a labeller's consistency (Dasgupta et al. 2023; Geiger et al. 2021; Desmond et al. 2021b). To help reduce these issues and improve the experience, processes like Active Labelling adopt LLMs with viable social reasoning capabilities to improve the accuracy (Beck et al. 2024), cost (Wang et al. 2021) and time to completion (Desmond et al. 2021b,a; Beck et al. 2024; Ashktorab et al. 2021) of labelling tasks. Yet the limits of such labelling assistants need to be carefully analysed.

Whilst there is an impressive library of literature suggesting that LLMs are approaching the ToM threshold, their capability differs greatly under different testing environments (Gandhi et al. 2023; Wang et al. 2024; Xie and Park 2023). In addition, there is a deficit in our understanding of LLMs' qualitative reasoning skills in social tasks (Han et al. 2024; Gassmann, McConville, and Edwards 2023). In particular, evaluations of LLMs' ability to reason about interpersonal influence in conversations are seemingly absent.

In this paper, we examine the capability of LLMs to detect and reason about influence in online interactions. Using a new corpus of 11,000 LLM labels, we explored the accuracy of nine state-of-the-art LLMs at identifying different categories of interaction and influence in 312 scenarios. In addition, we asked each LLM to explain its decision, and manually evaluated these explanations to identify patterns in each model's ability to comprehend the interpersonal influence expressed in an interaction, and justification of labelling decisions. Using LLM justifications, we were able to link some common failures of the models to features of the original interactions, highlighting specific contexts in which LLMs may struggle at social reasoning tasks. In summary, our research questions (RQ) are as follows:

- **RQ1:** To what extent can interpersonal influence in online interactions be identified by state-of-the-art large language models?
- **RQ2:** How reliably can state-of-the-art large language models *explain* their labelling decisions for social rea-

soning tasks?

- **RQ3:** Which online interaction contexts do large language models struggle to comprehend?

The remainder of this paper is organised as follows: An overview of surrounding research covering general/social reasoning and automated labelling. Our research methodology, covering the data collection, cleaning process and LLM deployment. Our findings for each proposed RQ. Finally, we conclude with some key observations.

Related Work

The challenge of inferring social relationships is fundamental to understanding human behaviour (Gandhi et al. 2023). Sap et al. (2019b), Shapira et al. (2023), and Apperly (2010) define Social (or Neural) ToM (SToM) as recognising the social behaviours, beliefs, emotions and mental state of others, ranging from simple conversations to complex business negotiations. With the advent of social models like ChatGPT (GPT 3.5 and 4 Turbo), debate on the threshold for SToM (Shapira et al. 2023) has increasingly focused on whether existing LLMs show true comprehension or the Eliza effect (in which human traits are ascribed to machines that only mimic comprehension (Weizenbaum 1976)).

There are few large-scale resources for evaluating SToM (Sap et al. 2019b), with benchmarks ranging in quality and difficulty (Shapira et al. 2023). SocialIQA is a social-reasoning benchmark, containing 38,000 multiple choice questions (Sap et al. 2019b) and whilst popular, it has been criticised for being ambiguous and inconsistent. Compared to human performance, SocialIQA shows models underperform by 20% in social reasoning questioning. To identify the factors restricting LLM capabilities, Shapira et al. (2023) discovered models overly rely upon shallow heuristic patterns. Solutions like instruction tuning can significantly improve LLM social reasoning, for example, SocialiteLlama (an instruction-tuned model) provides superior zero-shot accuracy compared to few-shot Llama2 (Dey et al. 2024).

One application for LLM reasoning is assisted labelling, to assign one or more labels to classify an example for a specific task (Desmond et al. 2021b). This process is time-consuming and expensive (exasperated by multiple annotators (Malik et al. 2024)) (Wang et al. 2021; Ashktorab et al. 2021), with some data being ambiguous or uninteresting to the labeller (Desmond et al. 2021a) leading to inconsistent explanations (Malik et al. 2024). The risk of mistaken labels being used as the ground truth for research can lead to inaccurate performance measurements or training data. Geiger et al. (2021) show that only 26.7% of research uses non-repeated datasets, making it increasingly important to discuss whether these frequently-used datasets are reliable and whether human labellers can benefit from LLM assistance (Zhu et al. 2023; Ashktorab et al. 2021).

Human labelling research shows that whilst human-in-the-loop labelling is most common for dataset creation, it is an imperfect process with humans showing personal belief bias (Dasgupta et al. 2023) and requiring additional/expensive safeguards for sensitive or confidential data. Desmond et al. (2021b) and Dasgupta et al. (2023) demonstrate that these difficulties are further exaggerated when dealing with multi-label tasks and unfamiliar, fictional, or abstract scenarios. LLM assistance in labelling can help reduce the cost and time of labelling, while improving accuracy (Beck et al. 2024; Desmond et al. 2021a,b; IBM 2023; Zhu et al. 2023; Zhou et al. 2024). Lai and Tan (2019) give a promising demonstration of this approach in labelling data for deception detection, finding that showing model-predicted labels helped improve the accuracy of human labelling. However, the generalisation of this approach to labelling for other social reasoning tasks remains under-evaluated.

Methodology

We collected a dataset of social media content from the platform Mastodon using the Mastodon API. The Mastodon platform was chosen due to its API remaining open and the platform not relying on recommendation algorithms to promote artificial instances of influence. Our dataset was collected using a three-hop system in which each user is at most three steps removed from a source user. The source users were identified through a search for hyperlink references to the news outlet *BBC News*, and the extended network is collected through replies to posts made by users included in the dataset at an earlier stage. Our dataset can therefore be considered an extended community made up of Mastodon users who interact (if indirectly) with BBC News content. Due to the *small world theory*, the resulting dataset consists of 1 million unique post-and-reply pairs from 150,000 Mastodon accounts.

Manual labelling was then applied with each interaction being given one of five possible labels regarding the form of interpersonal influence observed in the interaction. The dataset was reduced by identifying post-and-replies with clear examples of the influence labels, and in doing so, identifying instances where influence was present in the first interaction. Any unclear examples were skipped. The extraction of clearly labelled examples across the five categories of influence resulted in a robust Cohen's Kappa agreement of 0.963. This score was attained by both annotators after a cross-examination of the labels and their corresponding reasoning. The final dataset consisted of 312 scenarios covering the array of online influence categories, over 2800 LLM labelled entries across four shot categories, totalling a human and AI labelled dataset of 11,000 labels for cross-comparison. The corresponding distribution of human influence labels and the corresponding definitions were:

- **Low Influence (24%):** Indicates that the user has responded in a neutral manner providing minimum data or purely factual information.
- **Moderate Influence (24%):** Indicates the user has responded with an opinion that aligns or indicates a personal level of interest in the discussion topic and what has been stated.
- **High Influence (19%):** Indicates an obvious level of change in the person's opinions caused by the initial comment.
- **Controversial Influence (10%):** Indicates a polarized comment caused by contempt to the original posted con-

tent, this influence indicates a reaction rather than a change in thought.

- **Repeating Influence (21%):** Indicates an aligned viewpoint likely caused by a similarity in perspective before the conversation took place, this influence indicates support to a pre-existing stance held.

Category	LLM Influence Predicted Labels				
Standard	*Low*	*Mod*	*High*	*Rep.*	*Cont.*
High Active	*Low*	*Active*		*Rep.*	*Cont.*
Active	*Active*			*Rep.*	*Cont.*
Binary Active	*Active*			*Non-Active*	

Table 1: Combination of labels used in sets. Standard labels have the following distribution: Low (24%), Moderate (24%), High (19%), Repeating (21%), Controversial (10%)

LLMs were then asked to predict these labels for post-and-reply pairs. Annotators were provided with the same influence definitions and instructions as those provided to the LLMs. We did this to aid comparison between LLMs and human annotators on social media, natural language, and instruction comprehension. We varied the number of examples made available to each LLM within the prompt, ranging from no examples (0-shot) to 3 examples of each label type (3-shot; 15 total examples). In analysis of LLM performance, the five influence labels were organised into different analysis sets [See Table 1] to test whether LLMs are more capable of distinguishing between different category boundaries when identifying types of interpersonal influence in online exchanges.

We tested the performance of an array of state-of-the-art LLMs. The Open-AI API was used to deploy the *GPT-3.5 Turbo* and *GPT-4 Turbo* models in January 2024, we also used a private account on the IBM Watson X AI Platform to deploy: *Flan T5 XXL, Flan UL2, Granite 13b Instruct V1, Granite 13b Instruct V2, Llama 2 13b Chat, Llama 2 70b Chat*, and *MT0 XXL*. These models all used the same hyperparameter settings: Repetition Penalty (1.0), Maximum New Tokens (300), Minimum New Tokens (0), Stop Sequence ([END]), Decoding Method (Sampling), Temperature (0.7), Top P (1), Top K (50).

While we evaluated the accuracy of each model at predicting the correct label for an online interaction, we also wanted to understand *why* an LLM was making a decision, and to evaluate its reasoning. For this reason, our prompt asked models to explain their reasoning. We then labelled each explanation following the scheme given:

1. **Inadequate Detail:** Short or no explanation.
2. **Detailed:** Long explanation offering unique perspectives that build upon the label definition.
3. **Invalid Structure:** The output content is incorrectly tagged or inconsistently structured.
4. **Post Content Hallucination (PCH):** False content that has been quoted as occurring in the posts or tangent like behaviour.
5. **Author Confusion:** Misquoting or misunderstanding the context of the discussion topic because of user confusion.
6. **References Content:** Quotes post or reply content.
7. **Lacked Conviction:** The user's explanation contains more than one label option or a lack of confidence in the predicted label.
8. **Label Misalignment:** The predicted label and the explanation describe different labels.
9. **Correct for Incorrect Reasons (CFIR):** Whilst an answer's label is correct the conclusion is not.
10. **Convincing:** The content provides a convincing alternative or correct answer.
11. **Unreadable:** The content is unreadable.
12. **Misinterpreted Cultural Context (MCC):** Incorrect prediction because of a misunderstood cultural reference.
13. **Notably Valid Cultural Interpretation (NVCI):** Showing an excellent understanding of human and cultural behaviour.

This allows us to correlate the features of an explanation with the accuracy of a corresponding decision, which is achieved by taking the highest and lowest 10% of examples based on accuracy across shot answers. As model accuracy could also depend on features of the online interaction, we also labelled the post-and-reply pair for the presence of common topical or content features. These labels were defined by observations from annotators during the first round of labelling, and examined relationships between the post content features, LLM reasoning, and model accuracy. These labels included the following categories and definitions:

1. **Sarcasm:** Sarcasm towards a person or topic.
2. **Topic Anger:** Anger towards the discussed topic.
3. **Public Information:** Publicly available information.
4. **Private Information:** Personal anecdote or private information.
5. **Requesting Advice:** A user is requesting advice or situational help.
6. **Pop-Culture:** References to music, films, books etc.
7. **Political:** Political people or topics.
8. **Question:** Content includes a question.
9. **News Story:** A discussion around a news topic.
10. **Quotes:** Content quotes a source.
11. **Promotions:** Content is promoting an associated product or service.
12. **Educational:** Content is educational in nature.
13. **Inspirational:** Content has an inspirational narrative.
14. **Personal Presentation:** A user is presenting a piece of work or an activity they are associated with.
15. **Interactive:** Content comments directly to other people's input.

Specification		Standard Labels		Active Influence		Binary Active Influence		High Active Influence	
Model	Shot	Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1
Flan T5 XXL	0	0.32	0.27	0.66	0.40	0.69	0.55	0.41	0.34
Flan T5 XXL	1	0.32	0.26	0.68	0.39	0.71	0.56	0.40	0.32
Flan T5 XXL	2	0.33	0.28	0.69	0.40	0.73	0.59	0.49	0.37
Flan T5 XXL	3	0.33	0.29	0.70	0.41	0.72	0.55	0.46	0.36
Flan UL2	0	0.32	0.23	0.68	0.35	0.70	0.52	0.42	0.30
Flan UL2	1	0.29	0.20	0.68	0.33	0.69	0.46	0.37	0.26
Flan UL2	2	0.30	0.21	0.67	0.30	0.69	0.48	0.39	0.26
Flan UL2	3	0.35	0.27	0.70	0.34	0.71	0.49	0.40	0.28
GPT-3.5 Turbo	0	0.22	0.15	0.69	0.31	0.70	0.48	0.32	0.22
GPT-3.5 Turbo	1	0.30	0.21	0.69	0.36	0.72	0.54	0.37	0.28
GPT-3.5 Turbo	2	0.31	0.28	0.68	0.42	0.71	0.57	0.44	0.35
GPT-3.5 Turbo	3	0.27	0.24	0.68	0.41	0.73	0.60	0.40	0.32
GPT 4 Turbo	0	0.37	0.31	0.72	0.53	0.76	0.69	0.47	0.44
GPT 4 Turbo	**1**	**0.42**	0.40	**0.75**	0.59	0.79	0.72	0.54	0.52
GPT 4 Turbo	**2**	0.41	0.40	0.70	0.59	0.76	0.74	**0.56**	**0.53**
GPT 4 Turbo	**3**	0.41	**0.41**	0.74	**0.60**	**0.80**	**0.77**	0.53	0.51
Granite V1	0	0.22	0.15	0.62	0.26	0.63	0.42	0.37	0.21
Granite V1	1	0.27	0.25	0.55	0.42	0.63	0.60	0.41	0.36
Granite V1	2	0.25	0.23	0.49	0.34	0.63	0.61	0.35	0.30
Granite V1	3	0.21	0.19	0.46	0.35	0.59	0.58	0.33	0.26
Granite V2	0	0.21	0.14	0.67	0.29	0.68	0.44	0.42	0.21
Granite V2	1	0.29	0.23	0.62	0.38	0.66	0.55	0.41	0.32
Granite V2	2	0.22	0.20	0.60	0.37	0.65	0.56	0.37	0.28
Granite V2	3	0.26	0.22	0.59	0.39	0.64	0.55	0.41	0.32
Llama 2 13b	0	0.23	0.08	0.68	0.27	0.68	0.41	0.43	0.15
Llama 2 13b	1	0.25	0.19	0.65	0.35	0.69	0.54	0.39	0.28
Llama 2 13b	2	0.19	0.17	0.54	0.30	0.65	0.58	0.34	0.24
Llama 2 13b	3	0.17	0.18	0.41	0.32	0.57	0.55	0.27	0.25
Llama 2 70b	0	0.24	0.14	0.67	0.31	0.68	0.44	0.40	0.21
Llama 2 70b	1	0.27	0.23	0.70	0.43	0.73	0.58	0.40	0.33
Llama 2 70b	2	0.31	0.31	0.65	0.49	0.68	0.61	0.47	0.40
Llama 2 70b	3	0.33	0.33	0.61	0.48	0.66	0.61	0.41	0.39
MT0 XXL	0	0.13	0.17	0.30	0.25	0.52	0.52	0.18	0.20
MT0 XXL	1	0.21	0.21	0.46	0.37	0.58	0.57	0.24	0.24
MT0 XXL	2	0.20	0.19	0.49	0.33	0.64	0.62	0.33	0.25
MT0 XXL	3	0.19	0.18	0.38	0.28	0.54	0.53	0.25	0.22
Mean		0.276	0.233	0.618	0.381	0.675	0.56	0.394	0.308
Std Dev		0.07	0.076	0.109	0.091	0.064	0.082	0.08	0.091

Table 2: Table showing the capabilities of models across all shots predicting influence labels using accuracy (Acc) and Macro-F1 (M-F1).

Results

Interpersonal Influence Identification (RQ1)

The capabilities of each model vary across shots [Table 2], with little evidence that shot quantity affects accuracy [reinforces Malik et al. (2024); Ziems et al. (2023)]. Judged under the Standard set of labels, the highest model accuracy obtained was 0.42 (GPT 4 Turbo, 1-shot), whilst the worst model performed at 0.13 (MT0 XXL, 0-shot). The best-performing models were the Flan (consistently above 0.3) and GPT models. Whilst all models under-performed in high-resolution influence labelling (the Standard label set; mean accuracy 0.276), they performed better in lower resolutions (e.g., mean accuracy for the Active label set rose to 0.618). The mean increase of 0.056 between Active and Binary Active labelling accuracy also implies fewer issues differentiating repeating and controversial forms of influence from active influence of any intensity.

We also explore whether model architecture/size affects predictive performance (GPT 4 parameter sizes are based on reported speculation). As shown in Table 4, while the largest model was the most performant, we find no consistent relationship between model size and influence prediction, with significantly more weight being put on training data over model size. In fact, Flan models with only 13b-20b parameters have second highest Standard set accuracy, outperforming the larger Llama2 and GPT 3.5 models.

Explanation Label	GPT-4	Llama 70b	MT0	Granite V1	Granite V2	Flan UL2	Flan T5	Llama 13b	GPT-3.5
(L) Inadequate Detail	**0.00**	0.32	50.00	19.94	20.44	27.45	24.32	1.29	0.70
(H) Detailed	**26.71**	21.66	NA	0.55	0.27	1.31	5.46	21.94	26.48
(L) Invalid Structure	**0.00**	10.83	50.00	19.94	20.44	26.80	24.32	10.97	3.14
(L) PCH	**1.19**	7.96	NA	11.91	7.08	4.58	5.19	1.61	2.44
(L) Author Confusion	**1.19**	2.87	NA	4.99	3.54	1.96	1.91	3.23	3.14
(H) References Content	**26.71**	22.61	NA	16.62	19.62	14.71	13.93	22.90	25.78
(L) Lacked Conviction	1.78	3.18	NA	1.94	2.45	0.98	**1.09**	4.19	4.18
(L) Label Misalignment	6.23	**5.10**	NA	5.54	7.08	6.54	6.01	5.81	6.27
(L) CFIR	**0.30**	0.96	NA	2.77	1.09	3.59	2.46	0.65	0.35
(H) Convincing	**19.29**	14.97	NA	6.37	5.99	3.27	7.38	14.84	13.24
(L) Unreadable	**0.00**	**0.00**	NA	0.28	0.82	0.98	0.55	**0.00**	**0.00**
(L) MCC	2.97	4.14	NA	6.93	5.18	**1.96**	2.73	4.84	2.44
(H) NVCI	**13.65**	5.41	NA	2.22	5.99	5.88	4.64	7.74	11.85

Table 3: Table showing explanation content factor percentages across models. (L): Low values are more desirable; (H) Higher values are more desirable.

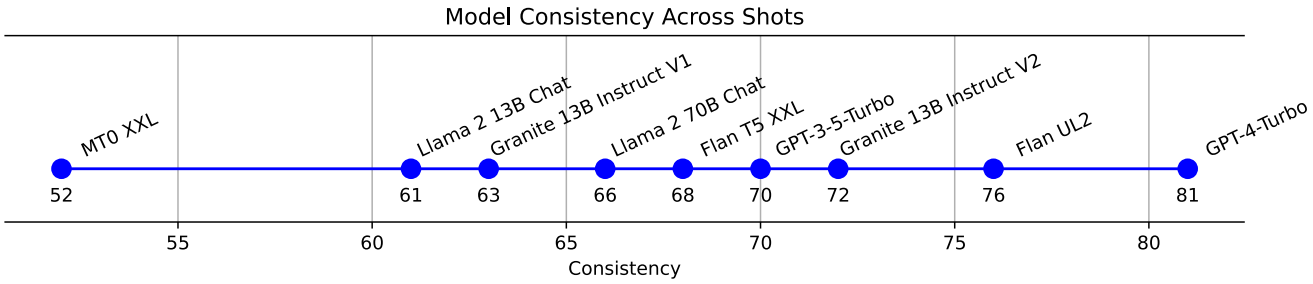

Figure 1: Chart showing the consistency of model labels across shots. Consistency is measured as the (percentage) occurrence of a dominant label for a piece of content across the four shot categories, which is then averaged across the entire dataset.

Model - Size	Std.	Active	Binary	High
GPT 4 - 100000B	**0.41**	**0.73**	**0.77**	**0.53**
GPT 3.5 - 175B	0.29	0.68	0.71	0.39
Llama 2 - 70B	0.29	0.66	0.68	0.41
Flan UL2 - 20B	0.31	0.68	0.69	0.4
MT0 XXL - 13B	0.19	0.42	0.56	0.25
Granite V1 - 13B	0.23	0.52	0.63	0.36
Granite V2 - 13B	0.24	0.61	0.65	0.41
Llama 2 - 13B	0.21	0.6	0.66	0.36
Flan T5 - 11B	0.33	0.69	0.71	0.44

Table 4: Table showing model parameter size to model prediction accuracy chart across categories.

Figure 1 presents how *consistently* a model produced the same label for a given post-and-reply pair across all of the 0-shot to 3-shot instruction sets. This measure gives us some understanding of how contingent upon examples a model's understanding of the influence labelling task might be. The most consistent model is GPT-4 Turbo (80%) followed by Flan and Granite, and whilst there is a decline, model averages do not fall below 50%, suggesting that models commonly have a dominant prediction across shots.

A consistent decision is not necessarily a correct one. However, Table 5 gives us insight into the accuracy of dominant labels and shows that, in practice, dominant labels are on average more accurate and have slightly less deviation across models in comparison to isolated shot predictions, with the largest improvement being in the Active label set (5.5%). Whilst these accuracy changes are minor, they suggest that the majority performance for each model can be improved with a collective strategy averaged across instruction sets involving different numbers of example decisions.

Interpersonal Influence Generated Explanations (RQ2)

We also review LLM explanations used to justify the assigned label. Using the categories showing model reasoning quality (shown in our Methodology), we labelled each explanation to assess the capability of each model's social reasoning. Variations in Table 3 show GPT-4 Turbo achieves the highest performance across explanation metrics. Due to the majority output of MT0 XXL being invalid, these scores were ignored when assessing these results. Regarding positive explanation characteristics, the *detailed* and *convincing* features are most often present in explanations authored

	Standard Labels		Active Influence		Binary Active Influence		High Active Influence	
Group	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Flan T5 XXL	0.33	0.27	0.69	0.41	0.71	0.55	0.45	0.36
Flan UL2	0.31	0.20	0.67	0.29	0.69	0.46	0.39	0.25
GPT-3.5 Turbo	0.29	0.24	0.71	0.40	0.73	0.57	0.41	0.32
GPT 4 Turbo	**0.42**	**0.39**	**0.75**	**0.61**	**0.80**	**0.75**	**0.54**	**0.52**
Granite 13b V1	0.25	0.23	0.60	0.41	0.69	0.63	0.40	0.33
Granite 13b V2	0.27	0.20	0.66	0.35	0.69	0.53	0.45	0.30
Llama 2 13b Chat	0.26	0.17	0.70	0.38	0.71	0.51	0.44	0.26
Llama 2 70b Chat	0.30	0.25	0.70	0.45	0.72	0.57	0.45	0.36
MT0 XXL	0.18	0.18	0.39	0.28	0.58	0.57	0.25	0.22
Mean	0.291	0.235	0.652	0.397	0.701	0.570	0.418	0.324
Standard Deviation	0.064	0.067	0.108	0.097	0.058	0.082	0.077	0.089
All Models	0.31	0.27	0.69	0.41	0.73	0.59	0.43	0.35

Table 5: Prediction accuracy using most common predicted label in each group.

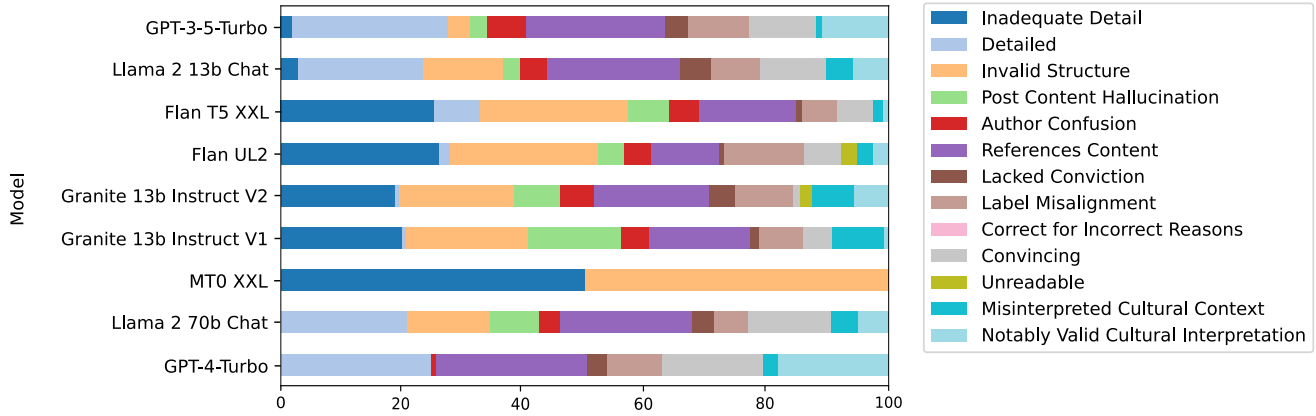

Figure 2: Chart showing percentages for each explanation factor for incorrectly predicted responses.

by the GPT and Llama models, whilst all models did well at *referencing content* of the online interaction. Regarding negative characteristics, post-content hallucinations (PCH) were most common in explanations produced by Granite and Llama 2 70B models, with Granite also being more likely to confuse the author with the discussion topic and misunderstand cultural contexts (MCC). We suspect that this is caused by the Granite models' training data focusing on ethically sourced and smaller datasets, demonstrating the importance of data over architecture (Xie and Park 2023; Dey et al. 2024). All models had a similar likelihood of label misalignment (5-7%), however, this was a rarer occurrence remaining below 10% per model.

Reviewing explanations specifically for cases where the label decision was incorrect (see Figure 2), we find the characteristics remain mostly the same as for correct predictions, with e.g., GPT-4 Turbo often presenting a correct interpretation of the post-and-reply exchange and a convincing argument for a label other than the correct one (GPT-3.5 Turbo and Llama models also did this to a lesser extent). The ability of these models to provide coherent justifications for an incorrect label speaks to both the utility and possible dangers of LLMs as annotators. Overall the scope of explanation quality was much broader than initially expected. Explanation characteristics like *Detailed*, *Convincing* and *NVCT* were often intercorrelated, whilst characteristics like *Author Confusion*, *Lacked Conviction*, *Label Misalignment*, *Unreadable* and *MCC* show little correlation with other characteristics or labelling performance, appearing to be independent issues.

Post and Reply Feature Analysis (RQ3)

By analysing the original features of the post-and-reply content, we can find content features that pose challenges for the models. Posted content annotations (found in our Methodology) were defined by observations from annotators during the first round of labelling, using these categories we identified common features of the online interactions that models were labelling, and compared the relationship between predictive accuracy and common features. For comparison purposes, we divide the models into clusters, using a K-Means clustering methodology using their labelling decisions. This methodology suggests three groups we devised based on model output clustering (before splinter groups oc-

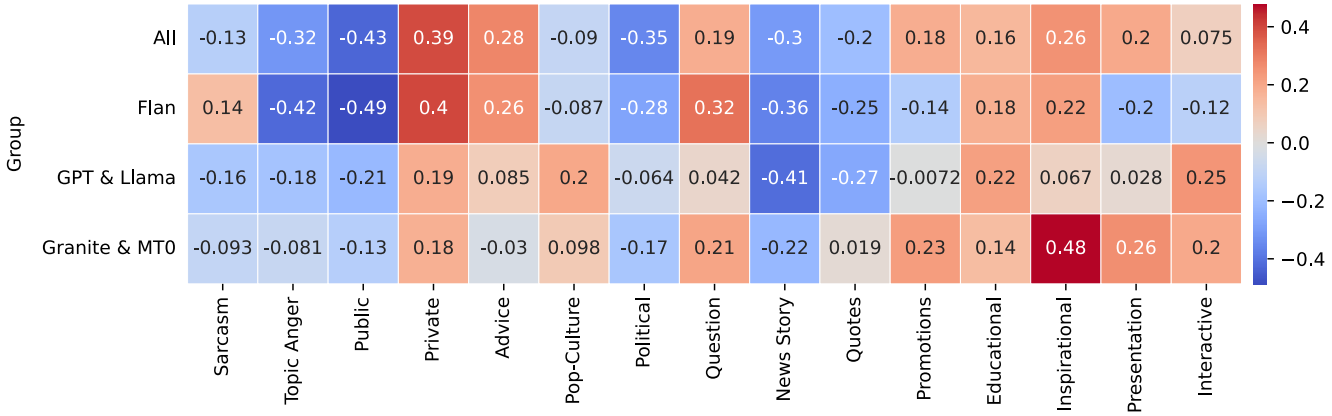

Figure 3: Heatmaps of model groups showing the correlation between social media post content features and the accuracy of each model group predicting influence.

cur): the *Flan* models, the *GPT and Llama* models, and the *Granite and MT0* models.

Figure 3 presents the correlations between accuracy and social media content features, both overall and per-cluster. Overall, we can see that references to publicly available information have the highest negative correlation to model accuracy, followed by references to political information, anger about the topic, and news stories. In general, this pattern suggests that the influence relationships in conversations related to negative comments on public or news forums are more likely to be misinterpreted by LLMs. In contrast, private or personal information alongside content that requested advice or was considered inspirational was positively correlated with model accuracy. This pattern holds true across other model groupings, with one differentiating factor being that the Granite and MT0 models' predictive performance was positively correlated with the presence of inspirational content.

We suggest that the negative correlations we observe are likely caused by models struggling to interpret newer unseen content/news (noted in (Shapira et al. 2023)). In addition, anger was seen to be harder for models to interpret, with models often confusing angry agreement on a topic with confrontational anger towards the other author. This challenge reflects a failure of models to demonstrate theory-of-mind: agents need to be identified separately from their environment (Zhou et al. 2023). Meanwhile, references to private information tended to make influence labelling significantly easier for the model, perhaps due to the less polarising nature of personal anecdotes and the higher likelihood that personal stories follow semantic patterns.

Conclusion

In summary, we present a novel experiment to test LLMs' capability for social reasoning and interpersonal influence detection. We focus on three research questions: the capability of LLMs identify forms of influence, their ability to explain their reasoning, and how their performance relates to common features of online content. Our results show that when asked to make high-resolution labelling decisions, LLMs fail, with all models falling below 45% accuracy in our evaluation. However, LLMs do perform significantly better when making decisions with lower resolution, with GPT-4 Turbo performing at 80% in binary choice tasks. We show that the Flan UL2 and GPT-4 Turbo models are the most consistent in behaviour across shots. Whilst GPT 3.5 and 4 Turbo provide accurate predictions in low-resolution environments, we see no consistent relationship between model size and accuracy in our research. The ability of LLMs to produce convincing explanations of their decisions is tested, and we discovered an array of issues, including a small but consistent rate for models producing explanations that contradict their labelling decision. More concerningly, models that produced convincing explanations for correct decisions were as capable of fabricating convincing explanations for incorrect decisions. Considering features of the original online interactions, we find that LLMs struggle more at providing an accurate prediction when confronted with public/news/angry content, whilst personal/inspirational content was easier for the models to label. In closing, LLMs provide some evidence of future influence labelling potential, and may someday provide substantial assistance in monitoring online social media content, monitoring social interactions, and labelling for social science purposes; however, existing LLMs require either careful handling or further refinement to overcome the issues we have discovered.

References

Apperly, I. 2010. *Mindreaders: The cognitive basis of "theory of mind"*. Psychology Press.

Ashktorab, Z.; Desmond, M.; Andres, J.; Muller, M.; Joshi, N. N.; Brachman, M.; Sharma, A.; Brimijoin, K.; Pan, Q.; Wolf, C. T.; et al. 2021. Ai-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–27.

Bar-Hillel, Y. 1960. The present status of automatic translation of languages. *Advances in Computers*, 1: 91–163.

Beck, N.; Killamsetty, K.; Kothawade, S.; and Iyer, R. 2024. Beyond Active Learning: Leveraging the Full Potential of Human Interaction via Auto-Labeling, Human Correction, and Human Verification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2881–2889.

Dasgupta, I.; Lampinen, A. K.; Chan, S. C. Y.; Sheahan, H. R.; Creswell, A.; Kumaran, D.; McClelland, J. L.; and Hill, F. 2023. Language models show human-like content effects on reasoning tasks. arXiv:2207.07051.

Desmond, M.; Duesterwald, E.; Brimijoin, K.; Brachman, M.; and Pan, Q. 2021a. Semi-automated data labeling. In *NeurIPS 2020 Competition and Demonstration Track*, 156–169. PMLR.

Desmond, M.; Muller, M.; Ashktorab, Z.; Dugan, C.; Duesterwald, E.; Brimijoin, K.; Finegan-Dollak, C.; Brachman, M.; Sharma, A.; Joshi, N. N.; et al. 2021b. Increasing the speed and accuracy of data labeling through an ai assisted interface. In *26th International Conference on Intelligent User Interfaces*, 392–401.

Dey, G.; Ganesan, A. V.; Lal, Y. K.; Shah, M.; Sinha, S.; Matero, M.; Giorgi, S.; Kulkarni, V.; and Schwartz, H. A. 2024. SOCIALITE-LLAMA: An Instruction-Tuned Model for Social Scientific Tasks. *arXiv preprint arXiv:2402.01980*.

Gandhi, K.; Fränken, J.-P.; Gerstenberg, T.; and Goodman, N. D. 2023. Understanding Social Reasoning in Language Models with Language Models. arXiv:2306.15448.

Gassmann, L.; McConville, R.; and Edwards, M. 2023. Predicting Interpersonal Influence from Conversational Features. In *2023 10th International Conference on Behavioural and Social Computing (BESC)*.

Geiger, R. S.; Cope, D.; Ip, J.; Lotosh, M.; Shah, A.; Weng, J.; and Tang, R. 2021. "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies*, 2(3): 795–827.

Han, S. J.; Ransom, K. J.; Perfors, A.; and Kemp, C. 2024. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83: 101155.

Hu, Z.; and Shu, T. 2023. Language Models, Agent Models, and World Models: The LAW for Machine Reasoning and Planning. *ArXiv*, abs/2312.05230.

IBM. 2023. Foundation models - IBM watsonx.ai.

Lai, V.; and Tan, C. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38.

Malik, U.; Bernard, S.; Pauchet, A.; Chatelain, C.; Picot-Clemente, R.; and Cortinovis, J. 2024. Pseudo-labeling with Large Language Models for Multi-label Emotion Classification of French Tweets. *IEEE Access*.

Manning, C. D. 2022. Human Language Understanding & Reasoning. *Daedalus*, 151(2): 127–138.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3027–3035.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019b. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4463–4473. Hong Kong, China: Association for Computational Linguistics.

Shapira, N.; Levy, M.; Alavi, S. H.; Zhou, X.; Choi, Y.; Goldberg, Y.; Sap, M.; and Shwartz, V. 2023. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. arXiv:2305.14763.

Shen, K.; and Kejriwal, M. 2023. An experimental study measuring the generalization of fine-tuned language representation models across commonsense reasoning benchmarks. *Expert Systems*, 40(5).

Wang, S.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4195–4205. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Wang, Y.; Chen, W.; Han, X.; Lin, X.; Zhao, H.; Liu, Y.; Zhai, B.; Yuan, J.; You, Q.; and Yang, H. 2024. Exploring the Reasoning Abilities of Multimodal Large Language Models (MLLMs): A Comprehensive Survey on Emerging Trends in Multimodal Reasoning. arXiv:2401.06805.

Weizenbaum, J. 1976. Computer power and human reason: From judgment to calculation.

Xie, B.; and Park, C. H. 2023. Multi-Modal Correlated Network with Emotional Reasoning Knowledge for Social Intelligence Question-Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3075–3081.

Zhou, P.; Madaan, A.; Potharaju, S. P.; Gupta, A.; McKee, K. R.; Holtzman, A.; Pujara, J.; Ren, X.; Mishra, S.; Nematzadeh, A.; Upadhyay, S.; and Faruqui, M. 2023. How FaR Are Large Language Models From Agents with Theory-of-Mind?

Zhou, Y.; Xu, P.; Wang, X.; Lu, X.; Gao, G.; and Ai, W. 2024. Emojis Decoded: Leveraging ChatGPT for Enhanced Understanding in Social Media Communications. arXiv:2402.01681.

Zhu, Y.; Zhang, P.; Haq, E.; Hui, P.; and Tyson, G. 2023. Can ChatGPT Reproduce Human-Generated Labels. *A Study of Social Computing Tasks*.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2023. Can Large Language Models Transform Computational Social Science?