Navigating Prompt Complexity for Zero-Shot Classification: A Study of Large Language Models in Computational Social Science

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, Xingyi Song

Department of Computer Science, The University of Sheffield
{y.mu, bpwu1, wthorne1, arobinson10, n.aletras, c.scarton, k.bontcheva, x.song}@sheffield.ac.uk

Abstract

Instruction-tuned Large Language Models (LLMs) have exhibited impressive language understanding and the capacity to generate responses that follow specific prompts. However, due to the computational demands associated with training these models, their applications often adopt a zero-shot setting. In this paper, we evaluate the zero-shot performance of two publicly accessible LLMs, ChatGPT and OpenAssistant, in the context of six Computational Social Science classification tasks, while also investigating the effects of various prompting strategies. Our experiments investigate the impact of prompt complexity, including the effect of incorporating label definitions into the prompt; use of synonyms for label names; and the influence of integrating past memories during foundation model training. The findings indicate that in a zero-shot setting, current LLMs are unable to match the performance of smaller, fine-tuned baseline transformer models (such as BERT-large). Additionally, we find that different prompting strategies can significantly affect classification accuracy, with variations in accuracy and F1 scores exceeding 10%.

Keywords: Large Language Model, Computational Social Science, Prompt Complexity

1. Introduction

Instruction fine-tuning (Ouyang et al., 2022) has facilitated transfer learning for Large Language Models (LLMs) to unseen tasks at scale. To leverage LLMs as versatile natural language processors, there is an immediate effort to ascertain their zero-shot performance on challenging tasks. Social media analysis is an active area of research with a number of complex, domain-specific tasks which can be utilised for harm reduction (Waseem et al., 2017) and preventing the spread of misinformation (Zubiaga et al., 2018). LLMs have great potential to assist with such computational social science (CSS) tasks, both in automatic data annotation and social media analysis (Kuzman et al., 2023; Reiss, 2023; Törnberg, 2023). Hence, it is important to understand the capabilities and limitations of the latest instruction fine-tuned LLMs for addressing such CSS tasks. In this paper, we are primarily focusing on answering the following research questions (RQ):

- **(RQ 1)** What level of zero-shot performance can LLMs achieve in social media classification tasks? How does zero-shot LLM performance compare against smaller state-of-the-art language models fine-tuned to the specific analysis task?
- **(RQ 2)** What are the most effective LLM prompt strategies for social media classification tasks in a zero-shot setting?
- **(RQ 3)** Was the pre-training corpus of the large model already inclusive of these datasets prior to the experiment (i.e., data leakage issues)?

To answer those research questions, we conduct a series of controlled experiments to investigate the zero-shot performance of two off-the-shelf instruction fine-tuned large language models using different prompting strategies. Namely, we experiment with GPT-3.5-turbo (GPT),[1] the most widely used proprietary instruction fine-tuned large language model; and OpenAssistant-LLaMA (LLaMA-OA) (Köpf et al., 2023), an open source LLM instruction fine-tuned based LLaMA (Touvron et al., 2023). We use six social media analysis NLP tasks to evaluate the classification performance of LLMs using different prompt complexity levels (including providing few-shots examples and publication information of benchmark datasets in the prompt). The findings are also compared against baselines employing standard techniques such as fine-tuning BERT.

It must be noted that the scope of this paper is on evaluating the performance of off-the-shelf, instruction fine-tuned language models on social media classification tasks, in a zero-shot setting. The evaluation of foundation language models without instruction fine-tuning is out of the scope of this paper.

Our main findings are:

[1] https://openai.com/blog/chatgpt

LREC-COLING 2024, pages 12074–12086

- **(i)** Task-specific fine-tuned models still generally tend to outperform LLMs in most zero-shot settings, even when the fully fine-tuned model (e.g., BERT-large model) is significantly smaller.
- **(ii)** Using prompting ensemble methods (e.g., on synonyms) can increase the performance and robustness of LLMs.
- **(iii)** Detailed and complex prompting strategies are not necessary.

2. Related Work

Both models evaluated in this work, GPT (also referred to as ChatGPT) and LLaMA-OA, have been trained using Reinforcement Learning with Human Feedback (RLHF) in conjunction with instruction tuning, as first explored in Ouyang et al. (2022). Instruction tuning is the fine-tuning of language models on NLP tasks rephrased as instructions and prior work has shown that it is an effective way of training LLMs to perform zero-shot on unseen tasks. (Wei et al., 2021; Sanh et al., 2021) Longpre et al. (2023) carried out a detailed ablation study on non-RLHF instruction tuning methods across the general NLP tasks in the Flan 2022 collection and found that T5 instruction tuned on the Flan performed surprisingly well on held-out tasks when compared to models directly fine-tuned on said task. Tuning with human feedback could be the next step in improving instruction tuning in this area.

Ziems et al. (2023) sets a roadmap for employing LLMs as data annotators by establishing prompting best practices and an evaluation of the zero-shot performance of 13 language models on 24 tasks in computational social sciences. In the financial domain, (Li et al., 2023) reveal that ChatGPT and GPT-4 outperform the performance of supervised models, which have been fine-tuned with domain-specific data, in several financial benchmarks.

To evaluate the zero-shot performance of ChatGPT for text classification, Kuzman et al. (2023) compares against a fine-tuned XLM-RoBERTa model for the task of automatic genre classification in English and Slovenian. They show that ChatGPT outperforms the baseline on unseen datasets and that there is no drop in performance when provided with Slovenian examples. Ganesan et al. (2023) use Facebook posts to classify user personality traits, based on openness, conscientiousness, extroversion, agreeableness, and neuroticism. They find that GPT-3 performs poorly on binary and worse yet on tertiary ranking for each trait.

LLMs have also been applied in mental health applications. Lamichhane (2023) evaluate ChatGPT's ability to classify stress, depression, and suicidal inclination from Reddit posts. Although ChatGPT significantly outperforms their baseline, the baseline consisted of a simple prediction of the majority class.

For toxicity detection, Wang and Chang (2022) analyse GPT-3's generative and discriminative zero-shot capabilities, finding that performance is only slightly better than a random baseline. However, the authors argue that the generative task allows for nuanced distinction of toxicity in the, somewhat subjective, binary setting.

Törnberg (2023) find that ChatGPT-4 outperforms non-expert annotators in identifying the political affiliation of Democratic or Republican party members based on their tweets during the 2020 US election. Wu et al. (2023) use ChatGPT to rank the conservatism of representatives in the 116th US Congress through a series of pairwise match ups, showing a high correlation with DW-NOMINATE scores.

As LLMs improve their performance on language generation tasks, the risk of misinformation and propaganda increases. Mitchell et al. (2023) propose DetectGPT, a perturbation-based zero-shot method for identifying machine-generated passages. (Su et al., 2023) further develop this approach with DetectLLM-LRR and -NPR, achieving improved efficiency and improved performance respectively.

Note that our work is distinct from previous research (Ziems et al., 2023); we evaluate Large Language Models (LLMs) on a different set of benchmarks and experiment with various prompt modification strategies, including replacing original labels with synonyms and incorporating arXiv paper titles.

3. Methodology

3.1. Prompting Strategies

Following the prompting approaches described by Child et al. (2019); Ziems et al. (2023), we develop prompts by (i) adding instructions after the context (e.g., task description) and (ii) using constraints (e.g., 'Only reply with Bragging' or 'Not Bragging.') at the end. We observe that using constraints can effectively avoid cases of model uncertainty (e.g., 'As an AI model, I cannot answer this question.') and guide models to generate the expected outputs.

For consistency, we use the same prompts for both GPT and LLaMA-OA. Examples of different prompt strategies are displayed in Table 1. To examine the zero-shot predictive performance of

Task	Basic
Bragging	**Basic Instruction** (i.e., Identify whether or not a tweet includes a bragging statement.) **+ Constraints** (i.e., Only reply (bragging) or (not bragging).) **+ Text** (e.g., Tweet: Come watch me and @USER face off in 2K best of 3 series #braggingrights @USER you next boiiii.)
Task	**Basic + T/L Desc**
Vaccine	**Basic Instruction** + **T/L Desc** Tweets that have been assigned to the class 'pro vaccine' express a positive opinion regarding the vaccination. Tweets belonging to the 'anti vaccine' class express a negative opinion towards COVID-19 vaccination. The 'neutral' class mainly includes news related to the development of vaccines, tweets that do not express a clear opinion, such as questions regarding the vaccine, informative tweets concerning vaccination. **+ Constraints + Text**
Task	**Few-sample**
Complaint	**Basic Instruction** **+ Few-samples** (e.g., (i) Complaint: @USER @USER give the timeline by which I'll receive my cashback which I should have received by 15th October 2017. (ii) Not Complaint: I just gave 5 stars to Nancy at @USER for the great service I received!) **+ Constraints + Text**
Task	**Memory Recall**
Hate Speech	**Basic Instruction** **+ arXiv Paper Title** (i.e., Recall this paper: Hateful symbols or hateful people? predictive features for hate speech detection on twitter.) **+ Constraints + Text**

Table 1: Prompt examples across different settings.

LLMs, we carry out a comprehensive set of experiments using four different prompting strategies.

Basic Instruction (Basic): We only provide a basic instruction without including detailed task and label descriptions. For example, for the bragging detection task, our prompt is: *'Identify whether or not a tweet includes a bragging statement.* + *Constraints* + *Text'.* Two possible configurations are tested, namely adding the prompt *before* or *after* the text respectively.

Task and Label Description (T/L Desc): Building upon the Basic Instruction Round, we provide additional information in the prompt by including task and label descriptions (see Table 1). Note that we use the labels and task descriptions detailed in the original papers on the respective datasets. The format of the prompts used for the Task and Label Description Round is: *'Basic Instruction* + *Task and Label Descriptions* + *Constraints* + *Text'*.

Few-sample Prompting (Few-sample): We also test a few-sample prompting strategy by adding one example selected from the training set for each label. The prompt designed for the few-sample experiments is: *'Basic Instruction* + *Few-shot Examples* + *Constraints* + *Text'*. Note that using few-sample as input is still a type of zero-shot setup, as we do not fine-tune the model.

Memory Recall (Recall): We observe that both GPT and LLaMA-OA can recall papers published before September 2021. Since arXiv papers are part of the training data of the LLMs, we also include the title of the source paper in the prompt when evaluating the model's zero-shot performance. For example, we include paper information by using this prompt: *'Recall this paper [Paper Title]* + *Basic Instruction* + *Constraints* + *Text'*. For such recall prompts, we only perform experiments on datasets published before September 2021. For reference, we examine the variations in performance across different checkpoints to assess whether instruction fine-tuning might influence the efficacy of the classification task.

3.2. Synonyms

LLMs might generate different outputs when using prompts which are semantically similar (e.g., synonyms[2]). To test the generalisability of LLMs, we substitute the names of each class with words that have the same or similar meaning. For example, we test the synonyms 'hateful', 'toxic', and 'abusive' to replace the original category 'offensive'. We also use two ensemble learning approaches to improve predictive performance by combining the outputs from all synonyms settings for each dataset:

- **Ensemble Majority:** We select the category that has been selected the most times across all synonym experiments.
- **Ensemble All Agreed:** We also experiment with a stricter setting that considers only model outputs that are in the same category (i.e., Complaint, Criticism, dissatisfaction, etc.) using all synonyms. For example, we consider the LLM that uses all synonyms

[2]Appropriate synonyms were selected by consulting https://www.thesaurus.com.

predicted as complaints, otherwise they are considered non-complaints. We only report this metric for datasets with binary classes.

4. Data

In order to ensure a comprehensive evaluation of LLM performance, we select six datasets that cover a wide range of computational social science tasks and different time spans. In particular, some of them were created before September 2021, while others were collected after the release of the LLMs used in this paper. All datasets are in English with manually annotated class labels. We detail dataset specifications and statistics in Table 2:

- **Complaint** This task aims to identify whether a tweet expresses a complaint, which is defined as 'a negative mismatch between reality and expectations in a particular situation' (e.g., customer complaints on Twitter) (Olshtain and Weinbach, 1987). We use a dataset developed by Preoţiuc-Pietro et al. (2019) consisting of 3,449 English tweets annotated with one of two categories, i.e., complaints or not complaints.
- **Vaccine Stance** This task aims to automatically predict the stance of tweets towards COVID-19 vaccination (Cotfas et al., 2021; Mu et al., 2023). The dataset developed by (Cotfas et al., 2021) provides 2,792 tweets belonging to one of three stance categories: pro-vaccine, anti-vaccine, or neutral.
- **Bragging** This task aims to classify whether a tweet is bragging or not bragging. We evaluate on a dataset developed by Jin et al. (2022) which contains 6,696 tweets labelled as either bragging or not bragging.
- **Rumour Stance** We use the RumorEval 2017 dataset which is developed by Derczynski et al. (2017). Here, we use the dataset for 4-way rumour stance classification, i.e., determining the stance of a reply towards a given source post (i.e. rumour) as either supporting, denying, questioning, or commenting.
- **Sarcasm** The sarcasm detection task is to identify whether a given tweet is intended to be sarcastic or not. We evaluate the task on the Semeval-2022 Task 6 dataset (Farha et al., 2022), which contains 4,868 tweets labelled as either sarcasm or non-sarcasm.
- **Hate Speech** The task of hate speech detection aims to study anti-social behaviours, e.g., racism and sexism in social media. We evaluate on a dataset developed by Waseem and Hovy (2016) with a binary classification setup, i.e., offensive or non-offensive.

5. Experimental Setup

5.1. Large Language Models

Our experiments are conducted using two publicly accessible large language models:

GPT-3.5-turbo (GPT) [3] is an enhanced version of the GPT-3 language model with instruction fine-tuning. GPT can be employed for a wide range of NLP tasks, including machine translation, common sense reasoning, and question answering. The experiments use the GPT model via the official OpenAI API.[4]

LLaMA-OA We employ the LLaMA-OA model developed by LAIONAI,[5] which fine-tunes the vanilla LLaMA (Touvron et al., 2023) 30B model using the OpenAssistant dataset (Köpf et al., 2023). Since the original LLaMA models are not allowed to be shared by individuals, LAIONAI could not release the weights for LLaMA-OA on huggingface but released xor (i.e., 'Exclusive Or') weights[6] applied to the original LLaMA weights and the check sum calculations performed to validate the conversion. In order to be able to run the experiments locally under hardware constraints, we applied 8-bit quantisation at model load time via BitsAndBytes (Dettmers et al., 2021) to decrease the inference memory requirements.

5.2. Baselines

The zero-shot classification performance of the two LLMs is compared against a weak Logistic Regression baseline and a strong fully fine-tuned BERT-large baseline:

Logistic Regression We represent the text using TF-IDF and consider tokens that appear more than 5 times.

BERT-large We fine-tune BERT-large[7] (Devlin et al., 2019) by adding a linear classifier on top of the 24-layer transformer blocks. The special token '[CLS]' is used as the representation of each text.

[3] https://platform.openai.com/docs/models/gpt-3-5

[4] https://platform.openai.com/docs/api-reference

[5] https://laion.ai/

[6] We use the OASST-SFT-6-LLAMA-30B version of the model. The xor weights can be found at: https://huggingface.co/OpenAssistant/oasst-sft-7-llama-30b-xor

[7] https://huggingface.co/bert-large-uncased

Dataset	# of Posts	Class (# of Posts)
Rumour Stance	5,568	Support (1,004) / Deny (415) / Query (464) / Comment (3,685)
Vaccine Stance	2,792	Pro Vaccine (991) / Anti Vaccine (791) / Neutral (1,010)
Complaint	3,449	Complaint (1,232) / Not Complaint (2,217)
Bragging	6,696	Bragging (781) / Not Bragging (5,915)
Sarcasm	4,868	Sarcasm (1,067) / Not Sarcasm (3,801)
Hate speech	16,907	Offensive (5,348) / Non-offensive (11,559)

Table 2: Dataset Specifications.

5.3. Data Splits

For each benchmark task, we divide the dataset into training (80%) and test (20%) sets using stratified random splits[8]. The training set is used for supervised fine-tuning, and is further sub-divided into a training and a validation subsets (in a 3:1 ratio) for hyperparameter tuning (e.g., early stopping) purposes. Subsequently, we evaluate the performance of the fine-tuned baselines and zero-shot LLMs on the 20% test set.

5.4. Evaluation Metrics

Performance results are reported using two evaluation metrics: 1) Accuracy which consists of a direct comparison between the model predictions and the ground truth label; and 2) F1-macro scores are reported for situations where accuracy may not provide an adequate representation of performance, particularly for certain imbalanced datasets, such as *Bragging* and *Rumour Stance*.

5.5. Hyper-parameters

During initial explorations, we observed that using a higher temperature (e.g., 0.8 for GPT and 2 for LLaMA-OA) results in inadequate classification performance, as it introduces more randomness in the model outputs. This suggests that higher temperature settings can cause the model outputs to be non-reproducible. Therefore in this study, we use a low temperature (i.e., 0.2)[9] for GPT to make the model more focused and deterministic.

For LLaMA-OA, we follow the 'precise hyper-parameter setup'[10] indicated in the OpenAssistant web interface, where the Temperature is 0.1, Top P is 0.95, Repetition Penalty is 1.2 and Top K is 50.

For BERT-large, we set the learning rate as 2e-5, the batch size as 16, and the maximum sequence length as 256. We run all baseline models three times with different random seeds and report average results. We fine-tune BERT-large on an Nvidia RTX Titan GPU with 24GB memory and run LLaMA-OA on an Nvidia A100 GPU with 40GB memory. The inference rates of LLaMA-OA and GPT are approximately 1,200 and 3,000 samples per hour respectively.

5.6. Reproducibility of LLM Output

As noted above, to ensure a consistent output, we utilise low temperature values of 0.2 and 0.1 for both GPT and LLaMA-OA. To evaluate the reproducibility of the models' output, we execute the basic prompt setting of the Complaint dataset five times for each language model. Our observations reveal that LLaMA-OA consistently generates identical outputs, whereas GPT achieves approximately 99% similarity in its outputs. Note that we consistently run LLaMA-OA on our own servers with identical hardware described in Section 5.5.

6. Results

The experimental results are shown in Table 3 and Table 4. Next we discuss them in relation to each of our three research questions.

(RQ 1) What level of zero-shot performance can LLMs achieve on social media classification tasks? How does zero-shot LLM performance compare against smaller state-of-the-art language models fine-tuned on the specific analysis task?

In general, LLMs (GPT and LLaMA-OA) with zero-shot settings are able to achieve better results than the simple supervised Logistic Regression model. However, the traditional smaller fine-tuned language model (BERT-large) still outperforms the two LLMs on the majority of the tasks (4 out of 6 tasks). Furthermore, we observe that GPT consistently outperforms LLaMA-OA across all prompt settings and tasks when considering only the F1-macro measure. However, our results show that the accuracy of LLaMA-OA is better than that of GPT on some imbalanced datasets, such as 'Bragging' and 'Sarcasm'. This may be due to LLaMA-OA defaulting to the neutral class (labels

[8]To generate class-stratified subsets, we employ a dataset split tool from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection

[9]https://platform.openai.com/docs/api-reference/chat/create

[10]https://open-assistant.io/dashboard

Model	Complaint		Vaccine Stance		Bragging	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Logistic Regression	81.4	79.7	72.8	73.1	88.6	58.8
BERT-large	89.4	88.6	**81.5**	**81.3**	**91.3**	**76.1**
GPT Basic After	84.9	84.1	65.5	65.8	81.1	62.7
GPT Basic Before	**89.7**	**88.7**	72.4	73.6	84.3	66.2
GPT T/L Desc	89.0	88.0	73.3	73.7	84.9	67.4
GPT Memory Recall	87.1	86.4	66.2	66.9	79.8	64.6
GPT Few-sample	85.6	85.2	68.2	69.4	77.3	61.8
LLaMA-OA Basic After	65.5	65.4	60.5	57.8	57.8	50.1
LLaMA-OA Basic Before	80.1	79.9	64.2	63.7	82.8	62.6
LLaMA-OA Basic (OAT 7)	83.9	83.4	66.4	65.9	64.1	42.0
LLaMA-OA T/L Desc	65.3	65.2	73.7	73.6	88.4	48.2
LLaMA-OA Memory Recall	82.6	82.1	64.2	63.8	88.1	46.8
LLaMA-OA Memory Recall (OA 7)	76.4	76.3	67.8	67.9	67.9	43.0
OA Few-sample	87.7	86.9	66.5	67.3	75.4	59.8
Model	**Rumor Stance**		**Sarcasm**		**Hate Speech**	
	Accuracy	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
Logistic Regression	68.5	40.9	76.1	53.5	83.2	79.2
BERT-large	**73.2**	**48.2**	**78.9**	58.4	**84.5**	**81.2**
GPT Basic After	53.0	36.2	74.3	**65.8**	72.9	77.0
GPT Basic Before	51.5	33.3	62.9	**59.7**	70.4	69.1
GPT T/L Desc	59.2	45.7	61.3	57.9	76.9	72.1
GPT Memory Recall	40.2	30.9	52.8	51.7	71.7	69.6
GPT Few-sample	40.8	30.6	68.9	64.9	74.8	71.8
LLaMA-OA Basic After	61.7	29.3	41.6	41.6	56.0	55.9
LLaMA-OA Basic Before	46.1	27.9	64.4	54.8	69.8	68.2
LLaMA-OA Basic (OAT 7)	63.1	35.4	61.4	38.8	58.1	58.1
LLaMA-OA T/L Desc	56.2	29.0	75.9	49.9	75.5	73.3
LLaMA-OA Memory Recall	52.4	34.6	78.1	43.9	55.4	55.4
LLaMA-OA Memory Recall (OA 7)	48.8	33.1	71.9	42.9	58.7	58.7
LLaMA-OA Few-sample	28.3	20.7	71.3	42.6	70.0	68.4

Table 3: LLMs zero-shot classification results across all prompt settings. All datasets are evaluated with accuracy and macro-F1 scores. Green highlighted cells denote prompt settings where zero-shot LLMs beat the supervised baseline (i.e., Bert-large model fine-tuned on the training set). **Bold text** denotes the best result per task. OA 7 denotes the 'OpenAssistant/oasst-sft-7-llama-30b-xor' model.

without any specific speech act, such as 'Not Bragging' and 'Not Sarcastic').

GPT achieves the best predictive performance on two speech act detection downstream tasks, namely *Complaint* (89.7 accuracy and 88.7 F1-macro) and Sarcasm (62.1 F1-macro). This suggests that LLMs can be employed as strong baseline models for zero-shot classification tasks.

With respect to prompts, when the results of T/L Desc and Memory Recall are compared against Basic Instruction, it is observed that using a more complex prompt (e.g., adding label and paper information) does not necessarily improve model performance and may even introduce additional noise, leading to a degradation in performance. This indicates that adding complexity to the prompt might lead to the LLM not fully focusing on the human instructions.

For speech act detection tasks such as *Complaint* and *Bragging*, the accuracy of LLMs exceeds 85%, indicating that LLMs can potentially be used for data annotation as a way to reduce human annotation costs. Standard data annotation tasks typically rely on at least two annotators in the first round, so one of them could be replaced by an LLM. According to the annotation details[11] of the vaccine stance task (Poddar et al., 2022), the agreement rate between the two annotators is approximately 62%.

(RQ 2) What are the most effective LLM prompt strategies for social media classification tasks in a zero-shot setting?

[11] https://github.com/sohampoddar26/covid-vax-stance/tree/main/dataset

Synonyms	GPT		LLaMA-OA	
	Accuracy	F1-macro	Accuracy	F1-macro
Task 1				
Complaint / not Complaint	**87.8**	**86.4**	80.1	79.9
Grievance / not Grievance	87.3	85.7	82.3	81.9
Criticism / not Criticism	80.4	77.9	76.7	76.4
Dissatisfaction / no Dissatisfaction	84.6	83.9	66.7	66.7
Discontent / no Discontent	80.7	80.0	55.2	54.2
Ensemble Majority	84.8	83.5	76.1	76.0
Ensemble All Agreed	86.8	85.1	**84.5**	**83.8**
Task 2				
Pro Vaccine / Anti Vaccine / Neutral	72.4	73.6	64.2	63.7
In Favour of the Vaccine / Against the Vaccine / Neutral	73.5	74.2	**64.4**	**63.9**
Positive Sentiment / Negative Sentiment / Neutral	70.8	70.8	58.9	52.5
Belief in vaccine / not Belief in Vaccine / Neutral	74.4	75.2	61.9	59.5
Positive Attitude to Vaccine / Negative Attitude / Neutral	72.3	72.3	63.7	61.3
Ensemble Majority	**74.7**	**75.4**	64.2	63.5
Task 3				
Bragging / not Bragging	84.3	**66.2**	82.8	62.6
Boasting / not Boasting	82.7	65.2	78.4	60.9
Showing off / not Showing off	78.8	62.9	**88.4**	56.3
Self-aggrandizing / not Self-aggrandizing	81.1	62.0	88.1	60.1
Excessively Proud / not Excessively Proud	75.2	58.0	77.9	58.1
Ensemble Majority	83.4	65.4	86.0	**63.9**
Ensemble All Agreed	**84.9**	64.4	88.1	59.8
Task 4				
Support / Deny / Query / Comment	**51.5**	**33.3**	46.1	27.9
Backing / Dismiss / Questioning / Comment	40.4	30.2	52.1	**43.8**
Support / Dismiss / Questioning / Comment	39.7	30.4	55.4	39.3
Ensemble	41.7	30.6	**55.5**	39.4
Task 5				
Sarcasm / not Sarcasm	62.9	59.7	64.4	**54.8**
Ironic / not Ironic	74.9	67.2	63.9	54.7
Insincere / Sincere	73.8	64.8	68.2	42.7
Disingenous / Genuine	77.8	61.9	56.8	49.3
Satire / not Satire	76.9	62.8	75.2	53.1
Ensemble Majority	74.9	**65.7**	70.5	53.9
Ensemble All Agreed	**80.1**	58.9	**76.9**	51.2
Task 6				
Offensive / Non-offensive	70.4	69.1	69.8	68.2
Toxic / not Toxic	64.1	63.5	70.7	67.8
Abusive / not Abusive	72.2	69.3	64.8	64.2
Hateful / not Hateful	73.9	71.2	**75.6**	**72.5**
Derogatory / not Derogatory	68.2	66.8	58.1	58.1
Ensemble Majority	71.4	69.7	73.6	71.1
Ensemble All Agreed	***75.1***	***71.6***	75.0	70.6

Table 4: LLMs zero-shot classification results using synonyms across all tasks. Green highlights are the original class names. Light grey highlighted cells denote where synonyms prompt settings beat the original label. **Bold text** denotes the best result per model per task.

Table 3 compares different prompt complexity, and shows that the simple prompt strategy works reasonably well. For GPT, adding task and label descriptions typically achieves better results, i.e. these prompts achieved the best results on 4 out of 6 datasets as compared to other GPT prompt strategies. On the other hand, LLaMA-OA achieves mixed results. On average, for LLaMA-OA, simple prompts outperform complex counterparts. This may happen because complex prompts add additional noise to the model. We also note that adding a few examples to the prompt actually damages classification performance, for both GPT and LLaMA-OA. We hypothesise that the longer prompt is affecting the model interpretation of instructions.

Datasets	# of Test Set	Average # of Errors	# of Unanimous Errors
Complaint	690	89	43
Vaxx Stance	559	145	82
Bragging	1,340	201	160
Rumor Stance	1,114	557	475
Sarcasm	974	194	58
Hate Speech	3,380	845	302

Table 5: We conduct further error analysis on the model outputs across all datasets. **# of Unanimous Error** denotes cases in which the LLM unanimously agrees on an incorrect answer while using different synonyms.

Task	Basic		T/L Desc	
	Tokens (Sum/Mean)	$	Tokens (Sum/Mean)	$
Rumour	35k/51	<0.1	82k/119	0.2
Vaccine	31k/127	<0.1	86k/45	0.2
Complaint	23k/33	<0.1	62k/91	0.1
Bragging	52k/76	0.1	96k/140	0.2
HateSpeech	62k/90	0.1	94k/137	0.2
Sarcasm	28k/41	<0.1	50k/86	0.1

Table 6: The cost of running GPT-3.5 for each task.

Table 4 shows all zero-shot results when synonyms are used in prompts for all six datasets. We observe that revising prompts with synonyms can substantially improve the zero-shot performance of LLaMA-OA, except for the *Bragging* dataset. It is worth noting that the Sarcasm dataset is the only one where the prompt using the original categories performs worse. This suggests that replacing original labels with synonyms allows the LLaMA-OA model to better understand the task requirements. The variation in the training example distribution for both GPT and LLaMA-OA could account for the observed behaviours of the models. For example, the LLaMA-OA model might be fine-tuned on a dataset like: '*[Text including offensive language]* + *[Category: Abusive]*'. Therefore, we believe that it is important to test similar words in place of the original labels when designing instructions as well as use ensemble methods.

(RQ 3) Was the pre-training corpus of the large model already inclusive of these datasets prior to the experiment (i.e., data leakage issues)? To answer this question, we test different prompting strategies (e.g., by asking about the authors and task details of each paper) to explore whether the LLMs have been exposed to the dataset beforehand. In Table 7, we present two examples of our testing approach by directly incorporating the titles of the RumourEval (Derczynski et al., 2017) and Sarcasm (Farha et al., 2022) datasets into the prompts. Considering that LLMs are capable of recalling task details when provided with the title of an arXiv paper (i.e., memory recall), we speculate that these LLMs might be trained on these source papers, incorporating some examples alongside their corresponding labels. However, due to the opaque nature of the training corpus utilised for these LLMs, it is uncertain to what extent these datasets were included in the training data.

7. Error Analysis

To better understand the limitations of LLMs, we conduct an error analysis focusing on shared errors across all synonym settings following (Ziems et al., 2023). We manually check these wrong predictions and observe that some unanimous errors (Ziems et al., 2023) (i.e., when the model agreed on an incorrect answer using different synonyms) are caused by incorrect or controversial ground truth labels. We summarise the number of wrong predictions from the synonyms experiments on GPT in Table 5.

On the other hand, we observe that LLaMA-OA often defaults to the majority category, such as 'not a bragging' and 'not sarcasm', which leads to higher accuracy but a lower macro-F1 measure. However, considering the high accuracy of LLM zero-shot classification performance, LLMs can still be utilised as data annotation tools (combined with human efforts) for NLP downstream tasks in CSS. We can utilise LLMs for data annotation and also to identify incorrect annotations.

8. Conclusion

This paper explored a number of prompting strategies for the application of Large Language Models (LLMs) in computational social science tasks. It presented a range of controlled experiments that establish the efficacy of different prompt strategies on six publicly available datasets. Our main findings are summarised as follows:

- Task-specific fine-tuned models generally tend to outperform LLMs in zero-shot settings.
- More detailed and complex prompts (e.g, by adding arXiv paper title and few-samples) do not necessarily enhance classification performance.
- The selection of specific words or phrases as the class label can considerably affect classification outcomes.

We therefore argue that developing prompts for zero-shot classification presents a significant challenge and recommend testing different prompt configurations before proceeding with experiments, while keeping in mind the time constraints[12] and financial costs associated with LLMs (see Table 6).

Limitations and Future Work

In this paper, we assess the zero-shot text classification performance of two popular LLMs (i.e., GPT and LLaMA-OA) on six downstream tasks in CSS. We acknowledge that further experiments on other fine-grained CSS tasks (such as Question&Answering and Named Entity Recognition) would be beneficial in future work. Currently, several state-of-the-art LLMs, such as GPT-4[13] and Mistral (Jiang et al., 2023), achieve better predictive performance than those evaluated in our work. Therefore, we plan to extend our experiments in the future to include more state-of-the-art LLMs of various sizes. Since our focus is primarily on out-of-the-box performance, we experiment with simple alterations of the prompts. Other research, e.g. Arora et al. (2022), has considered prompt aggregation as well as using LLMs to auto-generate prompts. We also do not explore advanced methods such as chain-of-thought prompting, which improves LM performance by encouraging it to output its intermediate reasoning steps (Wei et al., 2022; Suzgun et al., 2022).

We also tried to explore potential data leakage issues (Ziems et al., 2023) by testing various prompts to verify whether our test sets have been exposed to GPT and OA. However, due to the black box nature of the training datasets of these two LLMs, we are unable to confirm the presence of data leakage.

Ethics Statement

Our work has received ethical approval from the Ethics Committee of our university and complies with the research policies of Twitter. All datasets are obtained through the links provided in the respective research papers or by requesting them directly from the authors. Furthermore, we can confirm that the data has been fully anonymized before being fed to the LLMs for model inference.

Acknowledgements

This research is supported by a UKRI grant EP/W011212/1 ("XAIvsDisinfo: eXplainable AI Methods for Categorisation and Analysis of COVID-19 Vaccine Disinformation and Online Debates"[14]) and an EU Horizon 2020 grant (agreement no.871042) ("So-BigData++: European Integrated Infrastructure for Social Mining and BigData Analytics"[15]).

References

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel J Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Liviu-Adrian Cotfas, Camelia Delcea, Ioan Roxin, Corina Ioanăş, Dana Simona Gherai, and Federico Tajariol. 2021. The longest month: analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *Ieee Access*, 9:33203–33223.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

[12] https://platform.openai.com/docs/guides/rate-limits/overview

[13] https://openai.com/gpt-4

[14] https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/W011212/1

[15] http://www.sobigdata.eu

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in english and arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814.

Adithya V Ganesan, Yash Kumar Lal, August Håkan Nilsson, and H Andrew Schwartz. 2023. Systematic evaluation of gpt-3 for zero-shot personality estimation. *arXiv preprint arXiv:2306.01183*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mali Jin, Daniel Preoţiuc-Pietro, A Doğruöz, and Nikolaos Aletras. 2022. Automatic identification and classification of bragging in social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3945–3959.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Taja Kuzman, Nikola Ljubešić, and Igor Mozetič. 2023. Chatgpt: Beginning of an end of manual annotation? use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Yida Mu, Mali Jin, Charlie Grimshaw, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1052–1062.

Elite Olshtain and Liora Weinbach. 1987. 10. complaints: A study of speech act behavior among native and non-native speakers of hebrew. In *The pragmatic perspective*, page 195. John Benjamins.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Soham Poddar, Mainack Mondal, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2022. Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 782–793.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019.

Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

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

Yau-Shian Wang and Yingshan Chang. 2022. Toxicity detection with generative prompt-based inference. *arXiv preprint arXiv:2205.12390*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Patrick Y Wu, Joshua A Tucker, Jonathan Nagler, and Solomon Messing. 2023. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. *arXiv preprint arXiv:2303.12057*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

Appendix

Task & Prompt	
Rumour Stance	Do you know this paper: SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours?
Sarcasm	Do you know this paper: Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in English and Arabic?
Output	
GPT (Rumour)	Yes, I'm familiar with the paper titled SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. This paper is related to the SemEval-2017 workshop, specifically Task 8, which focused on rumour detection and veracity classification on social media platforms. Participants were provided with a dataset containing rumours along with associated tweets that either supported or denied the rumours. The task involved classifying the veracity of the rumours as true, false, or unverified, and also identifying the tweets that provided support or denial for the rumours.
LLaMA-OA (Rumour)	Yes, I am aware of the paper titled SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. This paper presents the results of the RumourEval task held as part of the 2017 edition of the SEMEVAL workshop. The task involved developing algorithms for determining the veracity of rumors and identifying sources that supported or refuted them using Twitter data.
GPT (Sarcasm)	I apologise for the confusion, but as an AI language model, my training only goes up until September 2021, and I do not have access to specific details about SemEval-2022 or any subsequent tasks. The last SemEval workshop I'm aware of is SemEval-2021.
LLaMA-OA (Sarcasm)	Semeval-2022 Task 6: ISARCASMEVAL is a subtask focusing specifically on the identification of ironic and sarcastic expressions in English and Arabic text. Irony and sarcasm are common forms of expression found in both written and spoken communication, but they pose significant challenges for automatic sentiment analysis due to their subtle nature and dependence on contextual cues.

Table 7: Ask LLMs if they can recall the original paper.

Tasks	Prompt with Task Description
Complaint	Complaining is a basic speech act used to express a negative mismatch between reality and expectations towards a state of affairs, product, organization or event. Key to the definition of complaints is the expression of the breach of expectations.
Vaccine Stance	Pro-vaccine tweets express a positive opinion regarding vaccination. Anti-vaccine tweets express a negative opinion towards COVID-19 vaccination. Neutral includes news related to vaccine development, questions about the vaccine, or informative tweets concerning vaccination without a clear opinion.
Rumour Stance	Support: the author of the response supports the veracity of the rumour. Deny: the author of the response denies the veracity of the rumour. Query: the author of the response asks for additional evidence in relation to the veracity of the rumour. Comment: the author of the response makes their own comment without a clear contribution to assessing the veracity of the rumour.
Hate Speech	A tweet is offensive if it: 1. uses a sexist or racial slur. 2. attacks a minority. 3. seeks to silence a minority. 4. criticizes a minority (without a well founded argument). 5. promotes, but does not directly use, hate speech or violent crime. 6. criticizes a minority and uses a straw man argument. 7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims. 8. shows support of problematic hash tags. E.g. "#BanIslam", "#whoriental", "#whitegenocide". 9. negatively stereotypes a minority. 10. defends xenophobia or sexism. 11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.
Sarcasm	Sarcasm is a form of verbal irony that occurs when there is a discrepancy between the literal and intended meanings of an utterance. Through this discrepancy, the speaker expresses their position towards a prior proposition, often in the form of surface contempt or derogation.
Bragging	Bragging is a speech act which explicitly or implicitly attributes credit to the speaker for some 'good' (possession, accomplishment, skill, etc.) which is positively valued by the speaker and the potential audience. As such, bragging includes announcements of accomplishments, and explicit positive evaluations of some aspect of self. A bragging statement should clearly express what the author is bragging about (i.e. the target of bragging).

Table 8: Task descriptions used for the prompting strategy **'Task and Label Description (T/L Desc)'**.

Tasks	Memory Recall Prompt
Complaint	Recall paper: (Automatically identifying complaints in social media). Identify whether a tweet is a customer complaint or a non-complaint. Only reply 'Complaint' or Non-complaint'.
Vaccine Stance	Recall paper: (The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics From Tweets in the Month Following the First Vaccine Announcement). Annotate a tweet into one of three stance categories: pro vaccine, anti vaccine, or neutral. Only reply the stance.
Rumour Stance	Recall paper: (SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours). Classify a tweet into one of four stance categories: support, deny, query, or comment. Only reply the stance.
Hate Speech	Recall paper: (Hateful symbols or hateful people? predictive features for hate speech detection on twitter.) Annotate whether a tweet is hateful or not hateful. Only reply hateful or not hateful.
Sarcasm	Recall paper: (Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in English and Arabic.) Annotate whether a tweet is sarcastic or not sarcastic. Only reply sarcastic or not sarcastic.
Bragging	Recall paper: (Automatic Identification and Classification of Bragging in Social Media) Identify whether or not a tweet includes a bragging statement. Only reply yes or no.

Table 9: Prompts for the Memory Recall strategy.

Tasks	Prompt with Few-sample
Complaint	Complaint: @USER @USER give the timeline by which I'll receive my cashback which I should have received by 15th October 2017. Not Complaint: I just gave 5 stars to Nancy at @USER for the great service I received!
Vaccine Stance	Pro Vaccine: This is very encouraging!! I've been hoping for a vaccine in 1Q21 so I can confidently travel to my favorite nephew's wedding in northern California in April. Anti Vaccine: This is why I have no faith that in Covid vaccines - Covid only harms those with compromised immune response, the very same people for whom vaccines don't seem to work (because their immune system isn't working properly). Neutral: Medical supplies across specialties are allocated under two budget items totalling RM4.29 billion in MOH's Specific Programmes in operating expenditure of Budget 2021; Covid-19 vaccines are also not listed for MOH under the separate Covid-19 Fund
Rumour Stance	Support: @USER @USER @USER @USER yeah i feel really sorry for them Deny: @USER I never called uber PT . Everyone is having a go at Uber but not PT ... We own it , we shouldn't have to pay in desperate times Query: @USER @USER Ironic since all the i witnesses say the officer was white . Now it is the black officer Darren Wilson who shot ? ? Comment: @USER @USER Uber is covering the cost of all rides , Uber is still paying drivers higher fares to encourage them to do pickups.
Hate Speech	Hateful: @USER Tell it to the 120 million Africans that Islam murdered. URL Not Hateful: @USER @USER doesn't look like I am.
Sarcasm	Sarcastic: I love days when Rob works short call and is only at the hospital for *checks watch* 13 hours. Not Sarcastic: I got stop putting on glitter flowers I'd like to ad red.
Bragging	Bragging: Come watch me and @USER face off in 2K best of 3 series #braggingrights @USER you next boiiii :flushed_face: :hot_face:. Not Bragging: I have completed survey on NaMo App.

Table 10: Examples used for the prompting strategy **'Few-sample'**.