Li, Jiahui; Papay, Sean; Klinger, Roman

Are Humans as Brittle as Large Language Models?

Date of secondary publication: 13.02.2026

Version of Record (Published Version), Conferenceobject

Persistent identifier: urn:nbn:de:bvb:473-irb-113162x

Primary publication

Li, J.; Papay, S.; Klinger, R. (2025): Are Humans as Brittle as Large Language Models?, in: K. Inui, S. Sakti, H. Wang, u. a. (Ed.), Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Mumbai, India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, pp. 2130–2155.

Are Humans as Brittle as Large Language Models?

Jiahui Li, Sean Papay and Roman Klinger
Fundamentals of Natural Language Processing, University of Bamberg, Germany
{jiahui.li,sean.papay,roman.klinger}@uni-bamberg.de

Abstract

The output of large language models (LLMs) is unstable, due both to non-determinism of the decoding process as well as to prompt brittleness. While the intrinsic non-determinism of LLM generation may mimic existing uncertainty in human annotations through distributional shifts in outputs, it is largely assumed, yet unexplored, that the prompt brittleness effect is unique to LLMs. This raises the question: do human annotators show similar sensitivity to prompt changes? If so, should prompt brittleness in LLMs be considered problematic? One may alternatively hypothesize that prompt brittleness correctly reflects human annotation variances. To fill this research gap, we systematically compare the effects of prompt modifications on LLMs and identical instruction modifications for human annotators, focusing on the question of whether humans are similarly sensitive to prompt perturbations. To study this, we prompt both humans and LLMs for a set of text classification tasks conditioned on prompt variations. Our findings indicate that both humans and LLMs exhibit increased brittleness in response to specific types of prompt modifications, particularly those involving the substitution of alternative label sets or label formats. However, the distribution of human judgments is less affected by typographical errors and reversed label order than that of LLMs.

1 Introduction

Large language models (LLMs) have shown impressive capabilities in automatic data annotation tasks (Tan et al., 2024). However, practical applications are often hindered by variability in model outputs, leading to inconsistent predictions (Zhao et al., 2024; Stureborg et al., 2024). This variability encompasses both the inherent non-determinism in probabilistic models (Song et al., 2025; Ouyang et al., 2025) as well as *prompt brittleness*, wherein minor changes in prompt phrasing lead to significant differences in outputs (Lu et al., 2022; Tang et al., 2024).

Of course, variability is also observed in human annotations. Annotators' individual experiences, sociodemographic backgrounds, and moral values shape how they interpret and label content. Moreover, the uncertainty in human annotation behavior may increase with the inclusion of additional annotators, posing challenges for explanation and identification (Pavlick and Kwiatkowski, 2019). Traditionally, such disagreement among annotators has been treated as noise or bias (Milkowski et al., 2021), often resolved through techniques like majority voting to produce a single "gold" label (Sabou et al., 2014; Anand et al., 2024).

Recent research suggests that variability in human annotations should not be seen as a "problem" but as a reflection of diverse human interpretations (Plank, 2022). For example, in an emotion labeling task, the text *'the dog ran towards me.'* may be labeled as joy or fear, depending on subjective preferences (Troiano et al., 2023).

As LLMs are trained to model human text, it is reasonable that their output distributions should somehow reflect the variability present in human annotations. However, while recent work has acknowledged the importance of human label variation, little attention has been paid to how this variation behaves across different prompt perturbations, particularly whether changes in annotator instructions affect the distribution of human responses analogously to LLM's prompt brittleness.

In this work, we draw a connection between human label variation and LLM predictions. Specifically, we investigate whether prompt brittleness – a model's sensitivity to small changes in task phrasing – is unique to LLMs, or if humans exhibit a comparable sensitivity to instruction variations. We propose a systematic method that formalizes prompt perturbation and investigates the distributional effects of such perturbations on both LLMs

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2130–2155

and human annotators. This method supports our exploration of the following research questions:

RQ1: How do distributional shifts of LLM outputs reveal prompt brittleness?

RQ2: Are humans susceptible to prompt brittleness in ways comparable to LLMs?

RQ3: Which types of prompt variants yield similar responses between humans and LLMs?

The experimental repository and results are available at `https://www.uni-bamberg.de/en/nlproc/projects/inprompt/`.

2 Related Work

In the following, we discuss related work on variability in LLM outputs and human annotations.

2.1 Variability in LLMs

We focus on two factors for variability in LLMs: prompt brittleness and model non-determinism.

2.1.1 Prompt Brittleness

LLMs exhibit high sensitivity to variations in a prompt, even when changes are minimal (Brown et al., 2020; Jiang et al., 2020; Gao et al., 2021). This phenomenon is commonly referred to as prompt brittleness. Prior studies have examined a variety of prompt perturbations that can lead to substantial changes in model outputs.

One prominent issue is position bias, where LLMs exhibit preferences for labels appearing in specific positions within a prompt. Altering the order of answer options can substantially affect model behavior (Lu et al., 2022; Qiang et al., 2024; Wang et al., 2024a,c). Similarly, the position of the set of possible answers as a whole affects the output (Zheng et al., 2024).

Lexical perturbations such as typos or synonym substitutions also affect the output. While some findings suggest that LLMs are relatively robust to typographical errors (Movva et al., 2024; Wang et al., 2024b), other work highlights brittleness in response to paraphrased instructions, including variations in verbs, nouns, and conjunctions (Mizrahi et al., 2024). Ceron et al. (2024) explore the effect of instruction styles, comparing personal versus impersonal phrasing.

Various approaches exist to mitigate the effect of prompt variations, such as guiding the inference process (Yugeswardeenoo et al., 2024; Lampinen et al., 2022) or by requesting explanations (Ye and Durrett, 2022; Mishra et al., 2022). Similarly, eliciting verbalized confidence estimates can reduce calibration errors in model outputs (Tian et al., 2023).

In this work, we will propose methods of prompt perturbations which build upon the theoretical foundations laid out in these works.

2.1.2 LLM Non-determinism

LLMs not only yield inconsistent outputs due to prompt brittleness, but also exhibit inherent non-determinism. Notably, even with a temperature of zero, variability persists when invoking LLM APIs such as ChatGPT (Ouyang et al., 2025). This stochastic behavior poses challenges for reproducibility in downstream applications.

Song et al. (2025) investigate the relationship between model size and output stability, finding that smaller LLMs may yield more consistent outputs when comparing the greedy decoding with different sampling strategies. To address non-determinism in practical (non-greedy) sampling scenarios, where LLMs produce a distribution of outputs for the same prompt, Hayes et al. (2025) propose probabilistic discoverable extraction.

In this work, we employ non-deterministic sampling-based decoding strategies for LLMs, treating variability not as noise but as an intrinsic property of LLM output, to be analyzed on the same footing as human response variations.

2.2 Variability in Humans

In the context of human annotation, variations between annotators are commonly treated as noise, often attributed to annotation errors (Zhang and de Marneffe, 2021). However, Plank (2022) argues that variability in human annotations can also arise from factors such as subjectivity, perspectivism, annotator disagreement, and the existence of multiple plausible views. Conversely, in survey responses, inter-participant variability is expected and desired in order to reflect the diversity of the surveyed population. However, such responses may also be sensitive to how questions are asked (Kalton and Schuman, 2018), exhibiting similarities to the prompt-brittleness phenomenon observed for LLMs.

2.2.1 Subjectivity & Annotator Disagreement

Subjective tasks, such as hate speech detection and emotion classification, are inherently influenced by the value systems, cognitive frameworks, and personal experiences of annotators (Sap et al., 2019; Aroyo and Welty, 2015; Milkowski et al., 2021). Recent work has introduced this concept in humans to LLMs, e.g., with socio-demographic prompting (Schäfer et al., 2025; Dayanik et al., 2022). Our approach considers the distribution of predictions by prompting LLMs a large number of times without anchoring responses to predefined demographic categories.

While subjectivity and perspectivism highlight the inherent variability in human interpretation, annotator disagreement also points to varying annotations which might not be of any benefit. Disagreement is therefore also observed in tasks considered objective, demanding a unified consensus for both model training and evaluation (Uma et al., 2022).

For example, Plank et al. (2014) demonstrate that disagreements in part-of-speech tagging stem from linguistically debatable cases rather than annotator mistakes. Similarly, Webber and Joshi (2012) analyze the challenges of human disagreement in discourse properties. Semantic annotation tasks are also prone to disagreement. Sommerauer et al. (2020) evaluate such cases using multiple quality metrics. De Marneffe (2012) argue that fact judgments should be treated as distributions.

In contrast to prior work that explores why annotators disagree, our study focuses on how both humans and LLMs vary in a distribution of their responses, considering the above cases as a whole.

2.2.2 Survey Response Bias

In social science studies, human responses often exhibit biases influenced by different survey designs, such as variations in wording, synonyms, and question formats used in market research (Brace, 2008; Cox, 1980). Also, the order of answer options in questionnaires can give rise to acquiescence and recency effects (McClendon, 1991).

Tjuatja et al. (2024) present the only study we are aware of which studies the relation between LLMs and human variations. They demonstrate that commercial LLMs exhibit sensitivity to prompt biases that elicit minimal change in human responses. However, their exploration of prompt modifications primarily relies on survey-style prompt variants and does not study human annotation variability. In contrast, our work systematically generalizes a broader and more diverse set of prompt reformulations, aiming to reflect the correlation of variability between LLMs and human annotation behavior.

3 Prompt Perturbation Methods

To study the relation between human and model "brittleness", we employ a systematic method for generating variations from a single base prompt. We hypothesize that prompt perturbations may be categorized according to their potential influence on human annotation decisions, which may reflect the variability of LLMs. Thus, in order to obtain a sufficient diversity of prompt variations, we explicitly construct prompt modifications of two categories: *neutral* prompt modifications, which we hypothesize should not significantly affect human responses, and *sensitive* prompt modifications, which we hypothesize should.

Within each category, we specify a number of aspects of the base prompt which could be mutated, and, for each aspect, define one or more concrete prompt variations. Tables 1 and 2 summarize the modification methods employed in both categories. For example, in the neutral category, one aspect, *LabelOrd*, reflects the internal order of the provided labels. For this aspect, we specify two concrete variants: *rev*, which reverses the base prompt's label order (mutating from ascending to descending order), and *shuff*, which permutes the labels randomly.

Sensitive changes might include the introduction of emotionally connotated wording in the prompt (referred to as *Emo*). For this aspect, appending the sentence "Respond quickly!" to the prompt corresponds to the variant *fast*. Appendix A.2 provides the full list of evaluated prompts in our experiments, developed by these methods.

Importantly, prompt variations can modify both model input and the label set, meaning model outputs and human annotations must be interpreted with respect to the prompt variation.

4 Experimental Settings

Our experiments involve collection of data from two sources: LLM predictions and human annotations. In order to compare LLM prompt brittleness to any potential instruction-conditioned variability in humans, we adopt an experimental methodology which maintains parallelism between these two data sources. Thus, we will first discuss our datasets and tasks, which are common to both, before detailing our specific settings for obtaining LLM predictions and human annotations.

Datasets and Tasks. For both humans and LLMs, we consider four tasks from three distinct English-language datasets: (1) offensiveness rating and (2) politeness rating from the POPQUORN dataset (Pei and Jurgens, 2023), (3) irony detection from

Name	Type	Description	Variant	Example
Base	Base Prompt	A single base prompt to apply modifications.	—	How do you rate ... given one of the labels 'a', 'b', 'c', 'd'?
Imper	Imperative expression	Reframes the prompt as an imperative.	pls: inserts the word 'please' for politeness.	**Please** rate ... given one of the labels 'a', 'b', 'c', 'd'**.**
LabelOrd	Label order	Alters the internal order of provided labels within the prompt.	1. rev: reverse the labels order. 2. shuff: shuffle the labels at random.	1. How do you rate ... given one of the labels **'d', 'c', 'b', 'a'**? 2. How do you rate ... given one of the labels **'a', 'd', 'b', 'c'**?
LabelPos	Label position	Specifies the location of the provided label set as a whole.	1. start: presents at the start of the prompt. 2. end: presents at the end of the prompt.	1. **Given one of the labels** 'a', 'b', 'c', 'd', how do you rate ...? 2. How do you rate ... given one of the labels 'a', 'b', 'c', 'd'?
Syns	Synonym substitution	Replaces lexical items with synonyms, applied to specific part-of-speech categories.	1. verb: verbs. 2. noun: nouns excluding labels. 3. prep: preposition. 4. co: coordinating conjunction.	1. How do you **evaluate** ... given one of the labels 'a', 'b', 'c', 'd'? 2. How do you rate ... given one of the **scales** 'a', 'b', 'c', 'd'? 3. How do you rate ... **according to** one of the labels 'a', 'b', 'c', 'd'? 4. How do you rate ... given one of the labels 'a', 'b', 'c', **or** 'd'?
Typo	Typographical errors	Introduces minor typographical errors into the prompt wording.	1. task: typos in the content of task description. 2. label: typos in the provided labels.	1. How do you **raet** ... given one of the labels 'a', 'b', 'c', 'd'? 2. How do you rate ... given one of the labels 'a', **'β'**, 'c', 'd'?
Cap	Capitalization	Capitalizes the words in the prompt.	1. task: key words of task content. 2. label: all the provided labels.	1. How do you rate ... given one of the **LABELS** 'a', 'b', 'c', 'd'? 2. How do you rate ... given one of the labels **'A', 'B', 'C', 'D'**?
PM	Punctuation mark	Operates on punctuation marks in the prompt.	1. remove: removes existing punctuation. 2. add: adds punctuation. 3. replace: replaces punctuation with equivalent alternatives.	1. How do you rate ... given one of the labels **a, b, c, d**? 2. How do you rate ... given one of the labels**:** 'a', 'b', 'c', 'd'? 3. How do you rate ... given one of the labels 'a'**;** 'b'**;** 'c'**;** 'd'?

Table 1: Prompt modification methods in the neutral category. Modifications are bold in the text for each example. Our hypothesis for these prompt variations is that human's show lower variation than LLMs.

the EPIC dataset (Frenda et al., 2023), and (4) emotion classification from the CROWD-ENVENT dataset (Troiano et al., 2023). These tasks were selected as all are known to involve some level of annotator subjectivity. For each dataset, we construct a base prompt, which clearly and succinctly requests an annotation for a provided instance according to the task's label set. We then construct a set of prompt variations, as described in Section 3. A full list of all prompt variations thus obtained is detailed in Appendix A.2.

LLMs. We select five LLMs capable of local deployment for evaluation: LLaMA-3.1-8B and LLaMA-3.3-70B (Patterson et al., 2022), Mixtral-8x7B (MistralAI, 2023), Falcon3-7B (Team, 2024), and Mistral-7B (Jiang et al., 2023).[1]

All models are accessed and executed using the HuggingFace library ecosystem. In order to investigate the distribution of LLM outputs, we enable stochastic decoding via sampling during generation. Specifically, the decoding configuration was selected to match typical deployments, with `top_p=0.9`, `temperature=0.6`, and `top_k=50`. These settings correspond to the default configuration used for the LLaMA-3.1-8B and LLaMA-3.3-70B models in the HuggingFace transformers library. The experiments are run on Nvidia A40 and Nvidia A100 GPUs, with a total estimation of 13,500 GPU hours for our study.

Model predictions are elicited in a zero-shot fash-

[1] `https://huggingface.co/`: `tiiuae/Falcon3-7B-Instruct`, `mistralai/Mixtral-8x7B-Instruct-v0.1`, `mistralai/Mistral-7B-Instruct-v0.3`, `meta-llama/Llama-3.3-70B-Instruct`, `meta-llama/Llama-3.1-8B-Instruct`.

Name	Type	Description	Variant	Example
AltLab	Alternative labels	Modifies the label set using synonymous or alternative mapping formats like Likert scales.	1. gran: changes the granularity of the label set. 2. int: changes the intensity of the label set. 3. keep: synonyms of labels.	1. How do you rate ... given one of the labels ‘a’, **‘a+’**, ‘b’, **‘b+’**, ‘c’, ‘d’? 2. How do you rate ... given one of the labels ‘a’, ‘b’, **‘d’**, **‘e’**? 3. How do you rate ... given one of the labels **‘a'’, ‘b'’, ‘c'’, ‘d'’**?
Def	Definition insertion	Adds definitions of the task content to the prompt.	1. task: explains the meaning of the task. 2. label: explains the meanings of labels. 3. both: explains the meanings of both 1. & 2.	1. How do you rate ... given one of the labels ‘a’, ‘b’, ‘c’, ‘d’? **... means...** 2. How do you rate ... given one of the labels? **‘a’:..., ‘b’:..., ‘c’:..., ‘d’:...**
Conf	Confidence statement	Request a confidence score with the answer.	—	How do you rate ... given one of the labels ‘a’, ‘b’, ‘c’, ‘d’? **Provide your answer alongside a confidence score.**
Exp	Explanation request	Asks for a justification.	—	How do you rate ... given one of the labels ‘a’, ‘b’, ‘c’, ‘d’? **Provide your answer with a justification.**
Emo	Emotional wording	Adds emotionally charged content to the prompt.	1. trust: requires trust of the answer. 2. warn: content warning. 3. care: requires care when making the answer. 4. fast: requires completing the task fast.	1. How do you rate ... given one of the labels ‘a’, ‘b’, ‘c’, ‘d’? **Trust your answer!** 2. How do you rate ... given one of the labels ‘a’, ‘b’, ‘c’, ‘d’? **The text may contain offensive words.** 3. How do you rate ... given one of the labels ‘a’, ‘b’, ‘c’, ‘d’? **Be careful with your answer.** 4. How do you rate ... given one of the labels ‘a’, ‘b’, ‘c’, ‘d’? **Respond fast!**

Table 2: Prompt modification methods in the *sensitive* category. Modifications are bold in the text for each example. We assume that humans are more affected by these variations than by those in the neutral category.

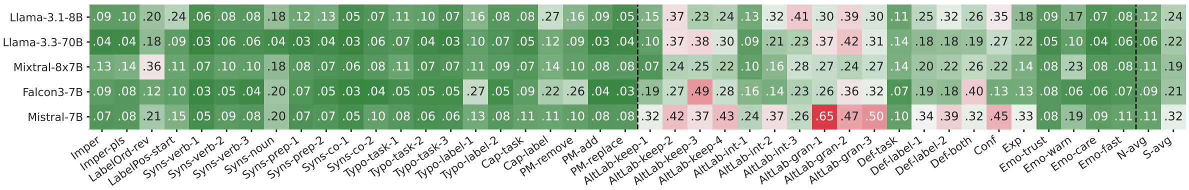
Task	Inst.	AnnNum
Offensiveness	1477	8.7
Politeness	3704	6.7
Irony	2987	5.0
Emotion	1161	5.0

Table 3: Statistics of the evaluated datasets. *Inst.*: Number of retained instances; *AnnNum*: Average number of annotators per instance.

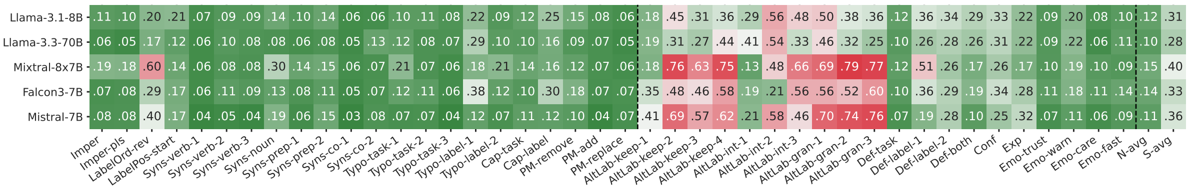
ion: For each instance, the model is presented with one prompt variation, followed by the instance text. The generated text is then parsed by attempting to extract one of the requested labels, with unparsable responses discarded. Across all tasks, for each (prompt variation, instance, LLM) triple, we elicit 100 responses in this manner. We discard instances for which we obtain less than 100 total valid predictions after 500 attempts. This leaves us with a total of 9329 retained instances across all tasks, for which the statistics are shown in Table 3.

Human Study. Due to the high cost of human annotation, we collect a subset of the judgments from our LLM study for human evaluation, focusing on two tasks: offensiveness rating and emotion classification. For each task, we sample approximately 10% of instances from the original POPQUORN and CROWD-ENVENT datasets, selecting an equal number of instances for each gold-standard label. We use ten prompts per task for the human study, selected so as to maximize the diversity of prompts tested. In addition to the base prompt, we select six variants which caused the most, medium and least extreme changes in LLM predictions from both the *sensitive* and *neutral* categories, as measured by Jensen–Shannon divergence (see Section 5). The other three variants are selected from the remaining types which present notable changes in instructing humans. Each prompt-instance pair is annotated by ten independent annotators recruited via the Prolific platform.[2] We employ only annotators based in the United States, with U.S. listed as their country of residence, country of birth, and nationality. All annotators are at least 18 years of age and demonstrated a high level of English proficiency, with English specified as their primary, first, and fluent language. We restrict the annotator approval rating to greater than 99%. Each annotator is presented with 22 instances including two attention

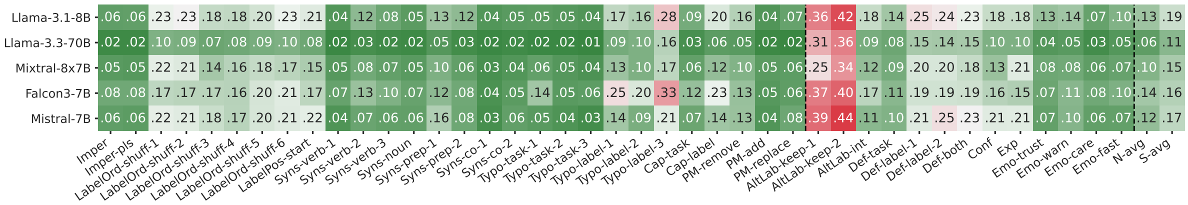
[2] https://www.prolific.com/



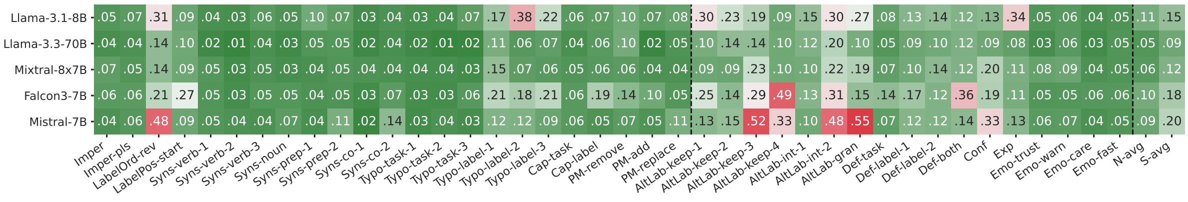
(a) Offensiveness rating.



(b) Politeness rating.

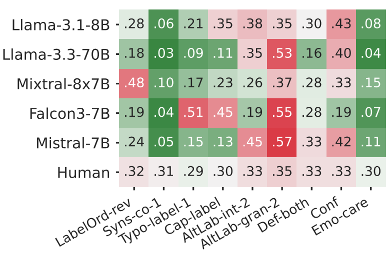


(c) Emotion classification.



(d) Irony detection.

Figure 1: Heatmaps showing distance scores between the distributions of LLM predictions with prompt variants and with the base prompt for four evaluated tasks in Table 3. Distance scores are calculated using Jensen–Shannon divergence (Dagan et al., 1997). The x-axis represents different types of prompt modifications defined in Section 3, with numeric indices representing multiple instances of the same type-variant pair. The y-axis lists the five LLMs evaluated in the study. Black dashed lines divide each heatmap into three parts: prompt modifications belonging to the neutral class (left), those from the sensitive class (middle), and the average distance scores for neutral and sensitive class modifications respectively (right). All prompts evaluated are provided in Appendix A.2.



(a) Offensiveness rating.

(b) Emotion classification.

Figure 2: Distance scores between the distributions of responses with prompt variants and the base prompt. The evaluated tasks are offensiveness rating (a) and emotion classification (b). The x-axis refers to the prompts with their modification types, and the y-axis refers to five LLMs and human samples. Distance scores are measured by Jensen–Shannon divergence (Dagan et al., 1997). Evaluated prompts can be found in Table 5 (a) and Table 8 (b).

checks, each preceded by a prompt variation acting as instructions. To mitigate potential biases arising from prior exposure to the same task, each annotator is assigned only one prompt variant per task. Each annotator is paid £9 per hour as suggested by Prolific. The median completion time for each survey is approximately 7–8 minutes, except for the *Exp* prompt variant, which requires about 16 minutes. The whole experiment costs £2,400, including 33.3% Prolific service fee.

5 Analysis

In this section, we present the results and analysis by answering three main research questions.

For both human annotators and LLMs, we quantify the distributional differences between predictions elicited by each prompt variant and those elicited by the base prompt. A full list of evaluated prompts is provided in Appendix A.2. For all the tasks, we employ the Jensen–Shannon divergence (Lin, 1991; Dagan et al., 1997) as a measure of distributional dissimilarity to display the human or LLM brittleness to prompt changes.

5.1 RQ1: How do distributional shifts of LLM outputs reveal prompt brittleness?

Figure 1 illustrates the distributional distances between various prompt variants and the corresponding base prompt across the four evaluated tasks. Across all tasks and five LLMs, we find that prompt changes categorized as *neutral* generally lead to smaller distributional shifts in model outputs compared to those categorized as *sensitive*. Interestingly, we observe that changes involving alternative label formulations, either semantically or in substitute mapping formats, exert the most important influence on model output distributions. Modifications including adding definitions and requiring justifications or a confidence score, which belong to the *sensitive* category, exhibit a moderate effect.

While these results broadly align with our hypothesis on the categorization of prompt types from the human-grounded aspect, some exceptions are noteworthy. Within the *neutral* category, changes to the internal order of labels (*LabelOrd*) induce substantial distributional shifts compared to changes to other prompt aspects. Within the *sensitive* category, introducing emotional language results in a relatively weaker impact than other modifications. Seemingly subtle changes, such as typos and capitalization of labels, yield measurable distributional shifts, highlighting LLM sensitivity to some surface variations, especially in labels.

Although these findings provide strong evidence of prompt brittleness, we observe that the LLaMA-3.3-70B model demonstrates greater robustness to prompt variations across all tasks, as indicated by its relatively lower average divergence scores. The Mixtral-8x7B model ranks second in robustness, except in the politeness rating task. This suggests that larger models tend to be more consistent to prompt perturbations than smaller ones.

To assess whether different LLMs exhibit similar patterns of brittleness across prompt modifications, we compute Spearman's rank correlation coefficients (Spearman, 2010) between pairs of models, based on the ranking of divergence scores induced by each prompt variant. Heatmaps of pairwise correlations are provided in Appendix A.1. For the emotion classification task, the models show high agreement in their ranking of prompt sensitivities, indicating a shared trend of distributional shifts. Although other tasks display lower rank correlations, the coefficients remain positive, suggesting a generally similar trend to different prompt modifications across models.

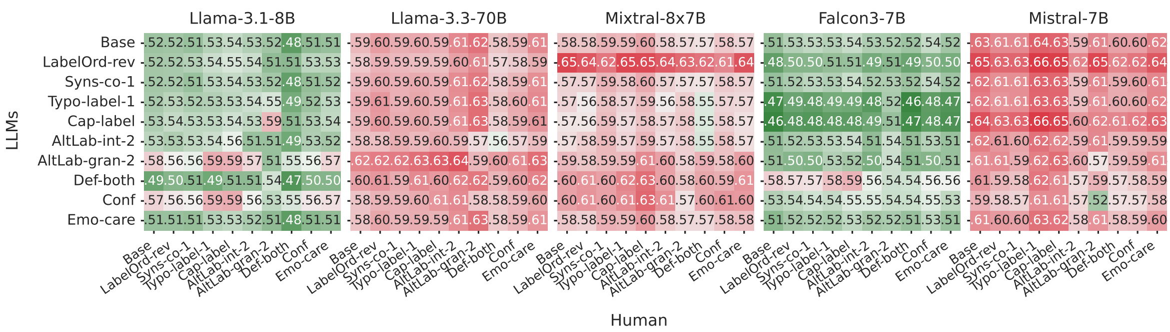
5.2 RQ2: Are humans susceptible to prompt brittleness in ways comparable to LLMs?

To investigate whether prompt brittleness also affects human annotators, we turn to our human annotation study, and compare the response distribution divergences across both LLMs and our human annotators. These results are presented in Figure 2.

Overall, we find that human divergences are higher for the same prompt, but seem to be less dramatically affected by the specifics of instruction variations than LLMs. Across both tasks, humans have an average divergence of 0.32, compared to an average LLM divergence of 0.22, but LLMs show a much larger standard deviation of 0.14, compared to 0.03 for humans. This is largely consistent with a state of affairs where human empirical distributions show greater instability, likely due to their smaller sample size, while showing less instruction-conditioned instability (i.e. brittleness) than LLMs.

However, while humans seem to be *less* brittle than LLMs, we still observe a clear brittleness effect in human annotations. In order to quantify human sensitivity to instruction variations, we employ aggregated χ^2-tests to compare each prompt variation to the base prompt, pooled across all instances which we assume to be independent. Table 4 sum-

(a) Offensiveness rating.



(b) Emotion classification.

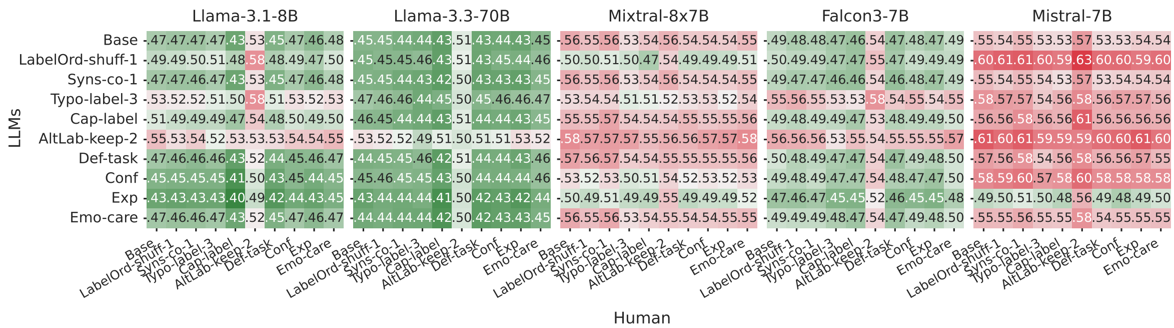


Figure 3: Distributional distance between LLM-generated and human annotations across five LLMs. The evaluated tasks are (a) offensiveness rating and (b) emotion classification. The x-axis and y-axis denote the prompt variants for human and LLMs respectively. The model name is displayed on the top of each subfigure. Evaluated prompts can be found in Table 5 for (a) and Table 8 for (b).

maries our findings: at a $p < 0.05$ level, we find statistically significant differences in human response distributions with *LabelOrd-rev*, *AltLab-gran-2*, *AltLab-int-2*, *Def-both*, *Conf*, and *Def-both* for the offensiveness rating task, and with *Typo-label-3*, *Emo-care*, *AltLab-keep-2* for the emotion classification task.

Interestingly, the prompt variations for which we find significant differences largely correspond to those which showed the largest distributional differences for LLM responses, with the sole exception of *Emo-care*, which affected human responses much more than it did LLMs. Moreover, although human annotation results remain consistent across both tasks, certain prompt variants (e.g., *Cap-label*, *Def-both*, and *Conf*) elicit markedly greater sensitivity in LLM performance on the offensiveness rating task than on the emotion classification task.

5.3 RQ3: Which types of prompt variants yield similar responses between humans and LLMs?

To explore the prompt variants that lead to similar response between humans and LLMs, we analyze the distributional distance between LLM-generated and human annotations, across prompt variants evaluated for the offensiveness rating and emotion classification tasks in our human study. Heatmaps presented in Figure 3a (offensiveness rating) and Figure 3b (emotion classification) illustrate the pairwise distributional differences between responses from each $(P_i(\text{LLM}_k), P_j(\text{human}))$ pair, where P_i and P_j denote specific prompt variants and LLM_k denotes one of the tested models.

For both tasks, we observe that LLaMA-3.1-8B and Falcon3-7B generally produces output distributions that are more aligned with human responses compared to other LLMs. This suggests that smaller LLMs may outperform larger models in approximating human behavior for some tasks.

Our analysis reveals that certain prompt modifications significantly increase the divergence in the distribution between LLM outputs and human annotations. In the offensiveness rating task, for example, the *LabelOrd-rev* prompt variant leads to notable deviations in outputs generated by Mixtral-8x7B and Mistral-7B, while the *AltLab-gran-2* variant causes pronounced discrepancies for the Llama models. Similarly, in the emotion classification

Task	Prompt Variant	χ^2
Offensiveness rating	LabelOrd-rev	607.2$*$
	Syns-co-1	(562.9)
	Typo-label-1	(511.3)
	Cap-label	(529.4)
	AltLab-int-2	621.4$**$
	AltLab-gran-2	704.2$***$
	Def-both	631.8$**$
	Conf	612.2$*$
	Emo-care	(530.9)
Emotion classification	LabelOrd-shuff1	(360.8)
	Syns-co-1	(425.4)
	Typo-label-3	469.9$**$
	Cap-label	(404.8)
	AltLab-keep-2	766.7$***$
	Def-task	(372.9)
	Conf	(390.4)
	Exp	(427.2)
	Emo-care	465.2$*$

Table 4: Aggregate χ^2-test results for human responses. In each case, we compare the human response distribution for a prompt variation against the response distribution for the base prompt, in order to test which prompt variations affected human responses in a statistically significant manner. Results reported are aggregated across all instances, responses for which are assumed to be independent. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, (): $p \geq 0.05$

task, the *AltLab-keep-2* and *Typo-label-3* variants tend to introduce higher divergence from human labels across most evaluated LLMs. Surprisingly, the *Exp* variant demonstrates that requesting an explanation during annotation enhances the alignment between LLMs and humans.

Conversely, we observe stronger alignment between LLM and human output distributions when both are presented with identical prompts, as opposed to mismatched prompt variants. This effect is particularly evident for the *AltLab-gran-2* and *Def-both* variants in the offensiveness rating task, especially with LLaMA-3.1-8B, LLaMA-3.3-70B, and Mistral-7B. Additionally, in the emotion classification task, the *Typo-label* and *Exp* variants yield relatively consistent alignment across almost all evaluated LLMs, which exhibits less divergence.

6 Conclusions and Future Work

In this paper, we investigate whether humans exhibit prompt brittleness, a sensitivity to prompt variations, in a manner comparable to LLMs. Our findings demonstrate that, similar to LLMs, human responses can be influenced by specific prompt modifications, although humans display greater robustness to certain types of changes. To explore this, we develop a systematic method of prompt perturbations grounded in the hypothesis that prompt modifications fall into two categories: those that affect the distribution of human annotations and those that do not.

Through human studies, we examine representative perturbations from each category. Notably, our analysis indicates that label substitutions induce comparable shifts in response distributions for both humans and LLMs. However, humans exhibit increased resilience to typographical errors and variations in label ordering where LLMs tend to struggle.

Furthermore, our results suggest that when LLMs and humans are presented with identical prompts, their output distributions are more aligned than when using mismatched prompts. This indicates that using consistent prompt formulations across both groups can facilitate better alignment in annotation tasks.

Not all our findings are consistent across models. While this variability may reflect differences in LLM training or model sensitivity to the stochastic decoding strategy, we offer practical insights for prompt engineering and emphasize the value of prompt-aware annotation.

Consequently, we advocate for future research to concentrate on objective tasks with lower inherent uncertainty than subjective tasks, to validate the generalizability of these results. Given the observed variability in LLM sensitivity to decoding strategies, further investigation into how decoding hyperparameters affect prompt brittleness is also warranted. Ultimately, our work underscores the necessity of accounting for prompt brittleness when designing human annotation tasks and interpreting LLM outputs, encouraging greater attention to prompt formulation in annotation contexts.

Limitations

While we explore all proposed categories of prompt variations, only a subset of samples within each category can be evaluated through human annotation due to practical constraints. Expanding the number of tasks and instances would further improve the robustness of our results. Among the various types of prompt perturbations investigated, some variants remain underexplored. For instance, in the case of the *LabelOrd* modification, we do not test all possible permutations of label orderings, which could limit the scope of our conclusions.

Additionally, the computation of distance scores using Jensen–Shannon divergence is sensitive to the size of the annotation distribution. In this study, the human annotations comprise only 10 samples, whereas the LLM-generated annotations comprise 100 samples. The relatively small human annotation set can result in higher variance in the distribution, which may increase the measured divergence compared to the larger LLM annotation set. Our comparisons between LLM-generated and human annotations could benefit from a more refined calibration strategy to address the discrepancy in the number of annotations per instance between the two sources.

Regarding the survey design, although we explicitly encourage annotators to read the prompt for every instance, full control over participant behavior is not feasible. It is possible that some annotators may have ignored the prompt after becoming familiar with the task structure.

Lastly, our experiments are conducted with only local LLMs. Introducing commercial LLM APIs could enhance the granularity of our study by contributing to real-world applications. This would, however, come at the cost of limited reproducability of the experiments.

Ethical Considerations

In our human study, all participants were informed that their responses, including demographic information, would be used for a scientific publication, and explicit consent was obtained. The collected data is anonymized to protect participants' privacy. Annotators who failed to pass one or more attention checks were excluded from the final results. We follow the instructions and suggestions of payment rate provided by the Prolific platform.

For the offensiveness rating task, participants were warned that the content they would be exposed to might include offensive or explicit language. In the emotion classification task involving prompts with intentional typographical errors, participants were informed afterwards that these typos were deliberately included for research purposes, to avoid confusion or misinterpretation. We do not change the intentional use for the datasets and models we refer to.

While one of the goals of this paper is to study the correlation of variability between LLMs predictions and human annotations, we acknowledge potential ethical concerns, particularly with regard to bias in both LLM outputs and human judgments. Nevertheless, we believe our findings provide valuable insights for future research in automatic annotation processes using LLMs.

ChatGPT (OpenAI, 2025) was used as a tool to improve code generation for figures and tables, as well as to assist with grammar and vocabulary in the text of some sections of this paper.

Acknowledgments

This work is funded by the project INPROMPT (Interactive Prompt Optimization with the Human in the Loop for Natural Language Understanding Model Development and Intervention, funded by the German Research Foundation, KL 2869/13–1, project number 521755488).

References

Abhishek Anand, Negar Mokhberian, Prathyusha Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2024. Don't blame the data, blame the model: Understanding noise and bias when learning from subjective annotations. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 102–113, St Julians, Malta. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Mag.*, 36(1):15–24.

I. Brace. 2008. *Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research*. Market research in practice series. Kogan Page.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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. *Transactions of the*

Association for Computational Linguistics, 12:1378–1400.

Eli P. Cox, III. 1980. The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17(4):407–422.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63, Madrid, Spain. Association for Computational Linguistics.

Erenay Dayanik, Ngoc Thang Vu, and Sebastian Padó. 2022. Bias identification and attribution in NLP models with regression and effect sizes. *Northern European Journal of Language Technology*, 8.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, Ilia Shumailov, Milad Nasr, Christopher A. Choquette-Choo, Katherine Lee, and A. Feder Cooper. 2025. Measuring memorization in language models via probabilistic extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9266–9291, Albuquerque, New Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Graham Kalton and Howard Schuman. 2018. The effect of the question on survey responses: A review. *Royal Statistical Society. Journal. Series A: General*, 145(1):42–57.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

McKee J. McClendon. 1991. Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research*, 20(1):60–103.

Piotr Milkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling, and Jan Kocon. 2021. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259, Online. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

MistralAI. 2023. Mixtral of experts.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in*

Natural Language Processing, pages 9048–9062, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2025. Chatgpt (gpt-4). https://chat.openai.com. Large language model.

Shuyin Ouyang, Jie M. Zhang, Mark Harman, and Meng Wang. 2025. An empirical study of the non-determinism of chatgpt in code generation. *ACM Trans. Softw. Eng. Methodol.*, 34(2).

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Preprint*, arXiv:2204.05149.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. 2024. Prompt perturbation consistency learning for robust language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1357–1370, St. Julian's, Malta. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wührl, Sabine Weber, and Roman Klinger. 2025. Which demographics do llms default to during annotation? *Preprint*, arXiv:2410.08820.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4195–4206, Albuquerque, New Mexico. Association for Computational Linguistics.

C Spearman. 2010. The proof and measurement of association between two things. *International Journal of Epidemiology*, 39(5):1137–1150.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *CoRR*, abs/2405.01724.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2024. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2327–2340, Mexico City, Mexico. Association for Computational Linguistics.

Falcon-LLM Team. 2024. The falcon 3 family of open models.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Rottger, and Barbara Plank. 2024b. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *First Conference on Language Modeling*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024c. “my answer is C”: First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Bonnie Webber and Aravind Joshi. 2012. Discourse structure and computation: Past, present and future. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54, Jeju Island, Korea. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA. Curran Associates Inc.

Dharunish Yugeswardeenoo, Kevin Zhu, and Sean O’Brien. 2024. Question-analysis prompting improves LLM performance in reasoning tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 402–413, Bangkok, Thailand. Association for Computational Linguistics.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

Xiutian Zhao, Ke Wang, and Wei Peng. 2024. Measuring the inconsistency of large language models in preferential ranking. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 171–176, Bangkok, Thailand. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

A Appendix

A.1 Heatmaps of SpearmanR Statistics Across LLMs

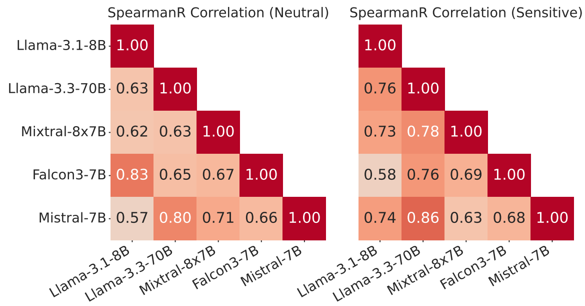
We compute Spearman’s rank correlation coefficients between pairs of models for the four evaluated tasks, based on the ranking of Jensen–Shannon divergence scores induced by each prompt variant. See Figure 4.

A.2 Examples for Prompt Modifications

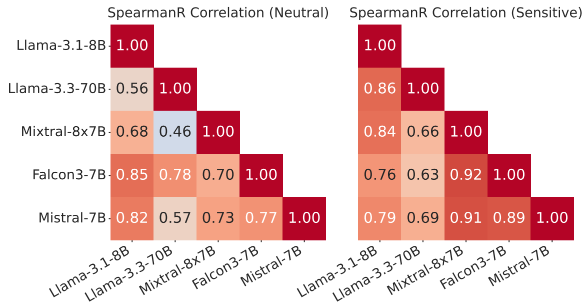
All prompt variations are derived from the ‘Base’ prompt, based on different types of modifications introduced in Section 3. The placeholder `{Inst.}` denotes the instruction that constrains the LLM generation. In our experiments, `{Inst.}` is instantiated with the sentence: "Provide only the label and omit any justification." We provide all prompts used across the four tasks in our study: Table 5 for offensiveness rating, Table 6 for politeness rating, Table 7 for irony detection, and Table 8 for emotion classification, along with their corresponding modification types.

A.3 Instructions of Survey for Human Study

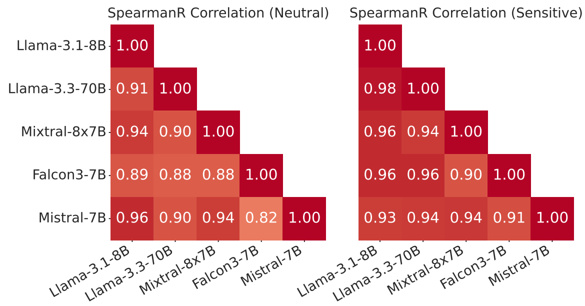
We provide the screenshots of the instruction page guiding our human study surveys. See Figure 5 for the offensiveness rating task and Figure 6 for the emotion classification task. An example of the annotation case for the emotion classification task, which involves a prompt with a typological error modification, is shown in Figure 7.



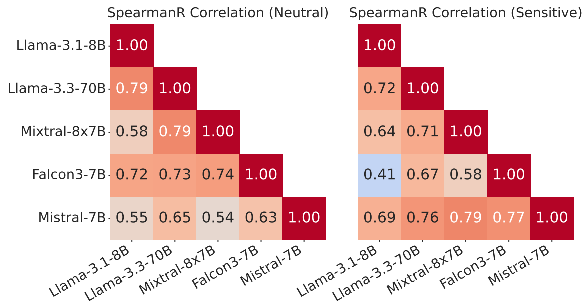
(a) Offensiveness rating.



(b) Politeness rating.



(c) Emotion classification.



(d) Irony detection.

Figure 4: Heatmaps showing Spearman's rank correlation coefficients of distance scores for prompt perturbations across five LLMs. The four evaluated tasks are (a) offensiveness rating, (b) politeness rating, (c) emotion classification, and (d) irony detection. We present the results for prompt perturbations in the neutral and sensitive categories introduced in Section 3. Labels of x-axis and y-axis both refer to the names of LLMs. The coefficient values have a range from -1 to 1, where 1 means perfect monotonic increasing correlation (ranks agree exactly), -1 means monotonic decreasing correlation (ranks are opposites), 0 means no monotonic relationship.

Table 5: Examples of prompt variants with different modification types used for the offensiveness rating task.

Type	Prompt
Base	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
Imper	Rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'. `{Inst.}`
Imper-pls	Please rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'. `{Inst.}`
LabelOrd-rev	How would you rate the offensiveness of the following text using one of the labels 'very offensive', 'offensive', 'moderately offensive', 'slightly offensive', or 'not offensive at all'? `{Inst.}`
LabelPos-start	Using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive', how would you rate the offensiveness of the following text? `{Inst.}`
Syns-verb-1	How do you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
Syns-verb-2	How would you evaluate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
Syns-verb-3	How would you judge the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
Syns-noun	How would you rate the offensiveness of the following text on the scale 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
Syns-prep-1	How would you rate the offensiveness of the following text according to the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
Syns-prep-2	How would you rate the offensiveness of the following text given the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
Syns-co-1	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', and 'very offensive'? `{Inst.}`
Syns-co-2	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', 'very offensive'? `{Inst.}`
Typo-task-1	How would you raet the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive','very offensive'? `{Inst.}`
Typo-task-2	How would you rate the offensivness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
Typo-task-3	How would you rate the offensiveness of the following text using one of the lables 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
Typo-label-1	How would you rate the offensiveness of the following text using one of the labels 'not offensiv at all', 'slightly offensiv', 'moderately offensiv', 'offensiv', or 'very offensiv'? `{Inst.}`
Typo-label-2	How would you rate the offensiveness of the following text using one of the labels 'not offensive at al', 'slightly offensive', 'moderatly offensive', 'offensive', or 'very offensive'? `{Inst.}`
Cap-task	How would you rate the OFFENSIVENESS of the following TEXT using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
Cap-label	How would you rate the offensiveness of the following text using one of the labels 'NOT OFFENSIVE AT ALL', 'SLIGHTLY OFFENSIVE', 'MODERATELY OFFENSIVE', 'OFFENSIVE', or 'VERY OFFENSIVE'? `{Inst.}`
PM-1	How would you rate the offensiveness of the following text using one of the labels not offensive at all, slightly offensive, moderately offensive, offensive, or very offensive? `{Inst.}`
PM-2	How would you rate the offensiveness of the following text using one of the labels: 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? `{Inst.}`
PM-3	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all'; 'slightly offensive'; 'moderately offensive'; 'offensive'; or 'very offensive'? `{Inst.}`
AltLab-keep-1	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'mildly offensive', 'fairly offensive', 'quite offensive', or 'very offensive'? `{Inst.}`
AltLab-keep-2	How would you rate the offensiveness of the following text on a Likert scale from 1-5, where 1 means 'not offensive at all' and 5 means 'very offensive'? `{Inst.}`
AltLab-keep-3	How would you rate the offensiveness of the following text on a Likert scale from 1-5, where 1 means 'not offensive at all', 2 means 'slightly offensive', 3 means 'moderately offensive', 4 means 'offensive', and 5 means 'very offensive'? `{Inst.}`
AltLab-keep-4	How would you rate the offensiveness of the following text on a Likert scale from 0-4, where 0 means 'not offensive at all' and 4 means 'very offensive'? `{Inst.}`
AltLab-int-1	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'very offensive', or 'extremely offensive'? `{Inst.}`
AltLab-int-2	How would you rate the offensiveness of the following text using one of the labels 'not offensive', 'mild', 'moderate', 'strong', or 'extreme'? `{Inst.}`
AltLab-int-3	How would you rate the offensiveness of the following text using one of the labels 'no offense', 'low', 'medium', 'high', or 'severe'? `{Inst.}`
AltLab-gran-1	How would you rate the offensiveness of the following text on a Likert scale from 1-4, where 1 means 'not offensive at all' and 4 means 'very offensive'? `{Inst.}`

Continued on next page

Table 5 – continued from previous page

Type	Prompt
AltLab-gran-2	How would you rate the offensiveness of the following text on a Likert scale from 1-7, where 1 means 'not offensive at all' and 7 means 'very offensive'? `{Inst.}`
AltLab-gran-3	How would you rate the offensiveness of the following text on a Likert scale from 1-10, where 1 means 'not offensive at all' and 10 means 'very offensive'? `{Inst.}`
Def-task	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? Offensiveness refers to how likely the text is to be perceived as inappropriate, insulting, or harmful in a general social context. Consider factors such as language, tone, and potential harm to individuals or groups. `{Inst.}`
Def-label-1	How would you rate the offensiveness of the following text using one of these labels? 'not offensive at all': The text contains no language or content that is likely to be considered offensive or inappropriate in most social and cultural contexts. 'slightly offensive': The text may contain mildly inappropriate language or implications, but it is unlikely to cause strong negative reactions. 'moderately offensive': The text includes content that some people might find offensive or inappropriate, but it is not explicitly harmful or strongly insulting. 'offensive': The text contains language or ideas that many people would consider inappropriate, offensive, or disrespectful. 'very offensive': The text includes highly inappropriate, harmful, or offensive content that is likely to cause strong negative reactions or be considered unacceptable in most contexts. `{Inst.}`
Def-label-2	How would you rate the offensiveness of the following text using one of these labels? 'not offensive at all': The text contains no language or content that could reasonably upset, insult, or demean any individual or group. It is neutral or positive in tone. 'slightly offensive': The text includes mildly insensitive or potentially provocative language, but the intent is ambiguous or the impact is minimal. 'moderately offensive': The text contains clearly disrespectful or derogatory content, though not extreme. It may offend some audiences but not others. 'offensive': The text uses overtly harmful language, slurs, or stereotypes likely to upset most readers. Intent may be hostile or careless. 'very offensive': The text includes severe slurs, threats, dehumanizing language, or hate speech that would widely be considered unacceptable. `{Inst.}`
Def-both	How would you rate the offensiveness of the following text using one of these labels? Offensiveness refers to how likely the text is to be perceived as inappropriate, insulting, or harmful in a general social context. Consider factors such as language, tone, and potential harm to individuals or groups. 'not offensive at all': The text contains no language or content that is likely to be considered offensive or inappropriate in most social and cultural contexts. 'slightly offensive': The text may contain mildly inappropriate language or implications, but it is unlikely to cause strong negative reactions. 'moderately offensive': The text includes content that some people might find offensive or inappropriate, but it is not explicitly harmful or strongly insulting. 'offensive': The text contains language or ideas that many people would consider inappropriate, offensive, or disrespectful. 'very offensive': The text includes highly inappropriate, harmful, or offensive content that is likely to cause strong negative reactions or be considered unacceptable in most contexts. `{Inst.}`
Conf	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? Provide the label along with a confidence score (1-5) of your rating, and omit any justification.
Exp	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? Provide the label along with a justification for your rating.
Emo-1	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? Important: Trust your gut reaction! `{Inst.}`
Emo-2	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? Warning: Some texts may be disturbing. `{Inst.}`
Emo-3	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? Please be careful with your rating. `{Inst.}`
Emo-4	How would you rate the offensiveness of the following text using one of the labels 'not offensive at all', 'slightly offensive', 'moderately offensive', 'offensive', or 'very offensive'? Important: Choose quickly and go with your first reaction! `{Inst.}`

Table 6: Examples of prompt variants with different modification types used for the politeness rating task.

Type	Prompt
Base	How would you rate the politeness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'? {Inst.}
Imper	Rate the politeness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'. {Inst.}
Imper-pls	Please rate the politeness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'. {Inst.}
LabelOrd-rev	How would you rate the politeness of the following text using one of the labels 'very polite', 'polite', 'moderately polite', 'slightly polite', or 'not polite at all'? {Inst.}
LabelPos-start	Using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite', how would you rate the politeness of the following text? {Inst.}
Syns-verb-1	How do you rate the politeness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'? {Inst.}
Syns-verb-2	How would you evaluate the politeness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'? {Inst.}
Syns-verb-3	How would you judge the politeness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'? {Inst.}
Syns-noun	How would you rate the politeness of the following text on the scale 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'? {Inst.}
Syns-prep-1	How would you rate the politeness of the following text according to the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'? {Inst.}
Syns-prep-2	How would you rate the politeness of the following text given the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'? {Inst.}
Syns-co-1	How would you rate the politeness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', and 'very polite'? {Inst.}
Syns-co-2	How would you rate the politeness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', 'very polite'? {Inst.}
Typo-task-1	How would you raet the politeness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite','very polite'? {Inst.}
Typo-task-2	How would you rate the politness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'? {Inst.}
Typo-task-3	How would you rate the politeness of the following text using one of the lables 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite' {Inst.}
Typo-label-1	How would you rate the politeness of the following text using one of the labels 'not polit at all', 'slightly polit', 'moderately polit', 'polit', or 'very polit'? {Inst.}
Typo-label-2	How would you rate the politeness of the following text using one of the labels 'not polite at al', 'slightly polite', 'moderatly polite', 'polite', or 'very polite' {Inst.}
Cap-task	How would you rate the POLITENESS of the following TEXT using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'? {Inst.}
Cap-label	How would you rate the politeness of the following text using one of the labels 'NOT POLITE AT ALL', 'SLIGHTLY POLITE', 'MODERATELY POLITE', 'POLITE', or 'VERY POLITE' {Inst.}
PM-1	How would you rate the politeness of the following text using one of the labels not polite at all, slightly polite, moderately polite, polite, or very polite? {Inst.}
PM-2	How would you rate the politeness of the following text using one of the labels: 'not polite at all', 'slightly polite', 'moderately polite', 'polite', or 'very polite'? {Inst.}
PM-3	How would you rate the politeness of the following text using one of the labels 'not polite at all'; 'slightly polite'; 'moderately polite'; 'polite'; or 'very polite'? {Inst.}
AltLab-keep-1	How would you rate the politeness of the following text using one of the labels 'not polite at all', 'mildly polite', 'fairly polite', 'quite polite', or 'very polite'? {Inst.}
AltLab-keep-2	How would you rate the politeness of the following text on a Likert scale from 1-5, where 1 means 'not polite at all' and 5 means 'very polite'? {Inst.}
AltLab-keep-3	How would you rate the politeness of the following text on a Likert scale from 1-5, where 1 means 'not polite at all', 2 means 'slightly polite', 3 means 'moderately polite', 4 means 'polite', and 5 means 'very polite'? {Inst.}
AltLab-keep-4	How would you rate the politeness of the following text on a Likert scale from 0-4, where 0 means 'not polite at all' and 4 means 'very polite'? {Inst.}
AltLab-int-1	How would you rate the politeness of the following text using one of the labels 'not polite at all', 'slightly polite', 'moderately polite', 'very polite', or 'extremely polite'? {Inst.}
AltLab-int-2	How would you rate the politeness of the following text using one of the labels 'not polite', 'mild', 'moderate', 'strong', or 'extreme'? {Inst.}
AltLab-int-3	How would you rate the politeness of the following text using one of the labels 'very impolite', 'impolite', 'neutral', 'polite', 'very polite' {Inst.}
AltLab-gran-1	How would you rate the politeness of the following text on a Likert scale from 1-4, where 1 means 'not polite at all' and 4 means 'very polite'? {Inst.}

Continued on next page

Table 6 – continued from previous page

Type	Prompt
AltLab-gran-2	How would you rate the politeness of the following text on a Likert scale from 1-7, where 1 means ‘not polite at all’ and 7 means ‘very polite’? `{Inst.}`
AltLab-gran-3	How would you rate the politeness of the following text on a Likert scale from 1-10, where 1 means ‘not polite at all’ and 10 means ‘very polite’? `{Inst.}`
Def-task	How would you rate the politeness of the following text using one of the labels ‘not polite at all’, ‘slightly polite’, ‘moderately polite’, ‘polite’, or ‘very polite’? Assessment is based on its choice of words, tone, and overall manner of expression. Consider whether the text demonstrates courtesy, respect, and appropriateness in communication. `{Inst.}`
Def-label-1	How would you rate the politeness of the following text using one of these labels? ‘not polite at all’: The text is rude, offensive, lacks any form of courtesy. It may include insults, aggressive language, a dismissive tone. ‘slightly polite’: The text is somewhat courteous but may contain blunt phrasing, minor rudeness, a lack of warmth. It is not outright offensive but could be perceived as unfriendly. ‘moderately polite’: The text maintains a neutral to mildly respectful tone. It avoids harsh language but does not go out of its way to be particularly warm courteous. ‘polite’: The text is respectful, considerate, and maintains a generally pleasant tone. It avoids any harsh dismissive language and follows social norms of politeness. ‘very polite’: The text is exceptionally courteous, warm, and respectful. It may include formal expressions of gratitude, apologies, other markers of strong politeness. `{Inst.}`
Def-label-2	How would you rate the politeness of the following text using one of these labels? ‘not polite at all’: The text lacks any politeness; it may be rude, blunt, or offensive. ‘slightly polite’: Minimal politeness. ‘moderately polite’: Clearly polite but neutral or formal; not overly warm. ‘polite’: Consistently courteous, with clear respect and positive tone. ‘very polite’: Exceptionally warm, respectful, or deferential; may include extra softening phrases. `{Inst.}`
Def-both	How would you rate the politeness of the following text using one of the labels ‘not polite at all’, ‘slightly polite’, ‘moderately polite’, ‘polite’, or ‘very polite’? Assessment is based on its choice of words, tone, and overall manner of expression. Consider whether the text demonstrates courtesy, respect, and appropriateness in communication. ‘not polite at all’: The text lacks any politeness; it may be rude, blunt, or offensive. ‘slightly polite’: Minimal politeness. ‘moderately polite’: Clearly polite but neutral or formal; not overly warm. ‘polite’: Consistently courteous, with clear respect and positive tone. ‘very polite’: Exceptionally warm, respectful, or deferential; may include extra softening phrases. `{Inst.}`
Conf	How would you rate the politeness of the following text using one of the labels ‘not polite at all’, ‘slightly polite’, ‘moderately polite’, ‘polite’, or ‘very polite’? Provide the label along with a confidence score (1-5) of your rating, and omit any justification.
Exp	How would you rate the politeness of the following text using one of the labels ‘not polite at all’, ‘slightly polite’, ‘moderately polite’, ‘polite’, or ‘very polite’? Provide the label along with a justification for your rating.
Emo-1	How would you rate the politeness of the following text using one of the labels ‘not polite at all’, ‘slightly polite’, ‘moderately polite’, ‘polite’, or ‘very polite’? Important: Trust your gut reaction! `{Inst.}`
Emo-2	How would you rate the politeness of the following text using one of the labels ‘not polite at all’, ‘slightly polite’, ‘moderately polite’, ‘polite’, or ‘very polite’? Warning: Some texts may be disturbing. `{Inst.}`
Emo-3	How would you rate the politeness of the following text using one of the labels ‘not polite at all’, ‘slightly polite’, ‘moderately polite’, ‘polite’, or ‘very polite’? Please be careful with your rating. `{Inst.}`
Emo-4	How would you rate the politeness of the following text using one of the labels ‘not polite at all’, ‘slightly polite’, ‘moderately polite’, ‘polite’, or ‘very polite’? Important: Choose quickly and go with your first reaction! `{Inst.}`

Table 7: Examples of prompt variants with different modification types used for the irony detection task.

Type	Prompt
Base	How would you classify the following reply to the given message using one of the labels 'ironic' or 'not ironic'? `{Inst.}`
Imper	Classify the following reply to the given message using one of the labels 'ironic' or 'not ironic'. `{Inst.}`
Imper-pls	Please classify the following reply to the given message using one of the labels 'ironic' or 'not ironic'. `{Inst.}`
LabelOrd-rev	How would you classify the following reply to the given message using one of the labels 'not ironic' or 'ironic'? `{Inst.}`
LabelPos-start	Using one of the labels 'ironic' or 'not ironic', how would you classify the following reply to the given message? `{Inst.}`
Syns-verb-1	How would you assess the following reply to the given message using one of the labels 'ironic' or 'not ironic'? `{Inst.}`
Syns-verb-2	How would you categorize the following reply to the given message using one of the labels 'ironic' or 'not ironic'? `{Inst.}`
Syns-verb-3	How would you identify the following reply to the given message using one of the labels 'ironic' or 'not ironic'? `{Inst.}`
Syns-noun	How would you classify the following reply to the given message using one of the categories 'ironic' or 'not ironic'? `{Inst.}`
Syns-prep-1	How would you classify the following reply to the given message according to the labels 'ironic' or 'not ironic'? `{Inst.}`
Syns-prep-2	How would you classify the following reply to the given message given the labels 'ironic' or 'not ironic'? `{Inst.}`
Syns-co-1	How would you classify the following reply to the given message using one of the labels 'ironic' and 'not ironic'? `{Inst.}`
Syns-co-2	How would you classify the following reply to the given message using one of the labels 'ironic', 'not ironic'? `{Inst.}`
Typo-task-1	How would you clasify the following reply to the given message using one of the labels 'ironic' or 'not ironic'? `{Inst.}`
Typo-task-2	How would you classify the following reply to the given mesage using one of the labels 'ironic' or 'not ironic'? `{Inst.}`
Typo-task-3	How would you classify the following reply to the given message using one of the lables 'ironic' or 'not ironic'? `{Inst.}`
Typo-label-1	How would you classify the following reply to the given message using one of the labels 'irnoic' or 'not irnoic'? `{Inst.}`
Typo-label-2	How would you classify the following reply to the given message using one of the labels 'ironc' or 'not ironc'? `{Inst.}`
Typo-label-3	How would you classify the following reply to the given message using one of the labels 'ironik' or 'not ironik'? `{Inst.}`
Cap-task	How would you CLASSIFY the following REPLY to the given MESSAGE using one of the labels 'ironic' or 'not ironic'? `{Inst.}`
Cap-label	How would you classify the following reply to the given message using one of the labels 'IRONIC' or 'NOT IRONIC'? `{Inst.}`
PM-1	How would you classify the following reply to the given message using one of the labels ironic or not ironic? `{Inst.}`
PM-2	How would you classify the following reply to the given message using one of the labels: 'ironic' or 'not ironic'? `{Inst.}`
PM-3	How would you classify the following reply to the given message using one of the labels 'ironic'; 'not ironic'? `{Inst.}`
AltLab-keep-1	How would you classify the following reply to the given message using one of the labels 'ironic' or 'literal'? `{Inst.}`
AltLab-keep-2	How would you classify the following reply to the given message using one of the labels 'ironic' or 'neutral'? `{Inst.}`
AltLab-keep-3	How would you classify the following reply to the given message using one of the binary labels 0 (ironic) or 1 (not ironic)? `{Inst.}`
AltLab-keep-4	How would you classify the following reply to the given message using one of the binary labels 1 (ironic) or 0 (not ironic)? `{Inst.}`
AltLab-int-1	How would you classify the following reply to the given message using one of the labels 'very ironic' or 'not ironic at all'? `{Inst.}`
AltLab-int-2	How would you classify the following reply to the given message using one of the labels 'negative' (ironic) or 'positive' (not ironic)? `{Inst.}`
AltLab-gran	How would you classify the following reply to the given message using one of the labels 'explicitly ironic', 'implicitly ironic', or 'not ironic'? `{Inst.}`

Continued on next page

Table 7 – continued from previous page

Type	Prompt
Def-task	How would you classify the following reply to the given message using one of the labels ‘ironic’ or ‘not ironic’? Irony implies a meaning that contrasts with its literal wording, often to express sarcasm, criticism, or humor. `{Inst.}`
Def-label-1	How would you classify the following reply to the given message using one of these labels? ‘ironic’: The reply expresses a meaning that is intentionally contrary to its surface content. ‘not ironic’: The reply is direct, sincere, and does not involve irony. `{Inst.}`
Def-label-2	How would you classify the following reply to the given message using one of these labels? ‘ironic’: The reply says one thing but means something else (usually the opposite). ‘not ironic’: The reply is straightforward and honest without hidden meaning—the words match the intent. `{Inst.}`
Def-both	How would you classify the following reply to the given message using one of these labels? Irony implies a meaning that contrasts with its literal wording, often to express sarcasm, criticism, or humor. ‘ironic’: The reply says one thing but means something else (usually the opposite). ‘not ironic’: The reply is straightforward and honest without hidden meaning—the words match the intent. `{Inst.}`
Conf	How would you classify the following reply to the given message using one of the labels ‘ironic’ or ‘not ironic’? Provide the label along with a confidence score (1-5) of your classification, and omit any justification.
Exp	How would you classify the following reply to the given message using one of the labels ‘ironic’ or ‘not ironic’? Provide the label along with a justification for your classification.
Emo-1	How would you classify the following reply to the given message using one of the labels ‘ironic’ or ‘not ironic’? Important: Trust your gut reaction! `{Inst.}`
Emo-2	How would you classify the following reply to the given message using one of the labels ‘ironic’ or ‘not ironic’? Warning: Some texts may be disturbing. `{Inst.}`
Emo-3	How would you classify the following reply to the given message using one of the labels ‘ironic’ or ‘not ironic’?Please be careful with your classification. `{Inst.}`
Emo-4	How would you classify the following reply to the given message using one of the labels ‘ironic’ or ‘not ironic’? Important: Choose quickly and go with your first reaction! `{Inst.}`

Table 8: Examples of prompt variants with different modification types used for the emotion classification task.

Type	Prompt
Base	How would you describe the emotion you infer from the following text using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’? `{Inst.}`
Imper	Describe the emotion you infer from the following text using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’. `{Inst.}`
Imper-pls	Please describe the emotion you infer from the following text using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’. `{Inst.}`
LabelOrd-shuff-1	How would you describe the emotion you infer from the following text using one of the labels ‘no-emotion’, ‘joy’, ‘surprise’, ‘pride’, ‘trust’, ‘relief’, ‘sadness’, ‘boredom’, ‘anger’, ‘fear’, ‘guilt’, ‘disgust’, or ‘shame’? `{Inst.}`
LabelOrd-shuff-2	How would you describe the emotion you infer from the following text using one of the labels ‘joy’, ‘surprise’, ‘pride’, ‘trust’, ‘relief’, ‘no-emotion’, ‘sadness’, ‘boredom’, ‘anger’, ‘fear’, ‘guilt’, ‘disgust’, or ‘shame’? `{Inst.}`
LabelOrd-shuff-3	How would you describe the emotion you infer from the following text using one of the labels ‘sadness’, ‘boredom’, ‘anger’, ‘fear’, ‘guilt’, ‘disgust’, ‘shame’, ‘no-emotion’, ‘joy’, ‘surprise’, ‘pride’, ‘trust’, or ‘relief’? `{Inst.}`
LabelOrd-shuff-4	How would you describe the emotion you infer from the following text using one of the labels ‘surprise’, ‘fear’, ‘shame’, ‘boredom’, ‘joy’, ‘guilt’, ‘sadness’, ‘trust’, ‘anger’, ‘relief’, ‘pride’, ‘no-emotion’, or ‘disgust’? `{Inst.}`
LabelOrd-shuff-5	How would you describe the emotion you infer from the following text using one of the labels ‘shame’, ‘sadness’, ‘trust’, ‘surprise’, ‘relief’, ‘guilt’, ‘joy’, ‘fear’, ‘disgust’, ‘pride’, ‘no-emotion’, ‘boredom’, or ‘anger’? `{Inst.}`
LabelOrd-shuff-6	How would you describe the emotion you infer from the following text using one of the labels ‘fear’, ‘shame’, ‘no-emotion’, ‘disgust’, ‘surprise’, ‘pride’, ‘relief’, ‘anger’, ‘boredom’, ‘joy’, ‘trust’, ‘sadness’, or ‘guilt’? `{Inst.}`
LabelPos-start	Using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’, how would you describe the emotion you infer from the following text? `{Inst.}`
Syns-verb-1	How would you characterize the emotion you infer from the following text using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’? `{Inst.}`
Syns-verb-2	How would you identify the emotion you infer from the following text using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’? `{Inst.}`
Syns-verb-3	How would you categorize the emotion you infer from the following text using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’? `{Inst.}`
Syns-noun	How would you describe the emotion you infer from the following text using one of the categories ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’? `{Inst.}`
Syns-prep-1	How would you describe the emotion you infer from the following text given the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’? `{Inst.}`
Syns-prep-2	How would you describe the emotion you infer from the following text according to the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’? `{Inst.}`
Syns-co-1	How would you describe the emotion you infer from the following text using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, and ‘no-emotion’? `{Inst.}`
Syns-co-2	How would you describe the emotion you infer from the following text using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, ‘no-emotion’? `{Inst.}`
Typo-task-1	How would you describe the emoton you infer from the following text using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’? `{Inst.}`
Typo-task-2	How would you describ the emotion you inferr from the following text using one of the labels ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’? `{Inst.}`
Typo-task-3	How would you describe the emotion you infer from the following text using one of the lables ‘pride’, ‘sadness’, ‘boredom’, ‘anger’, ‘joy’, ‘surprise’, ‘fear’, ‘guilt’, ‘disgust’, ‘trust’, ‘relief’, ‘shame’, or ‘no-emotion’? `{Inst.}`

Continued on next page

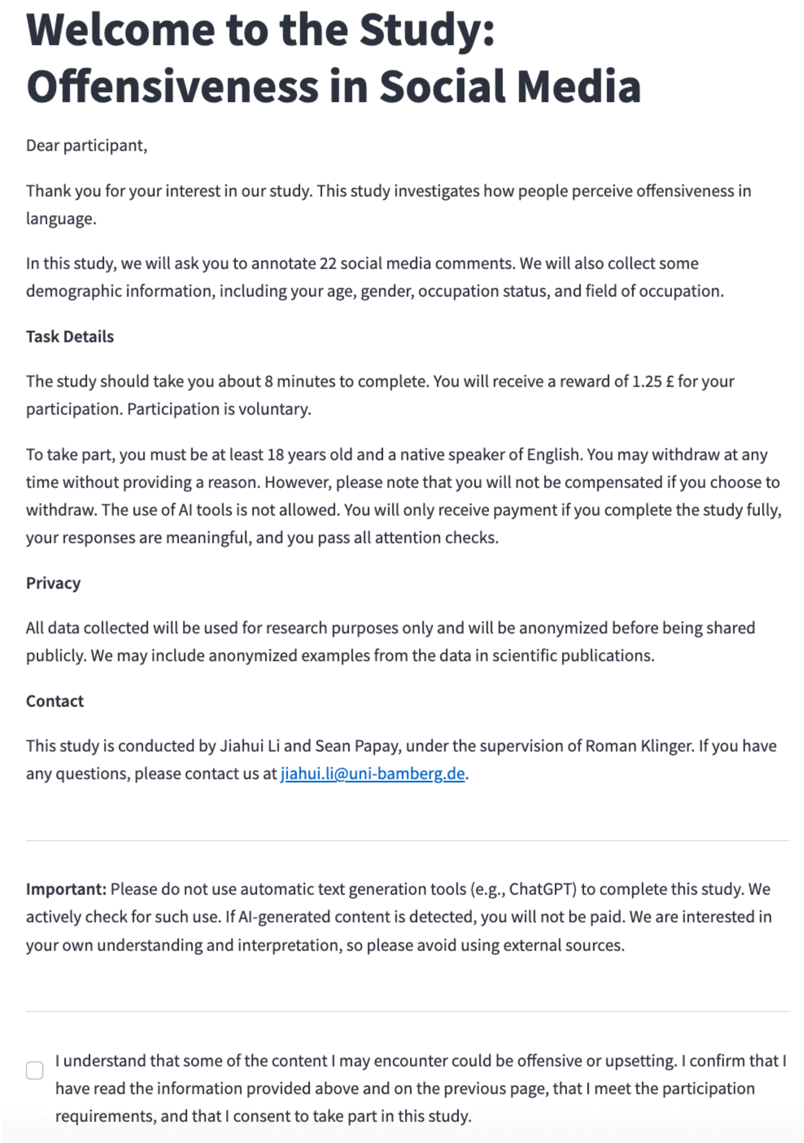
Table 8 – continued from previous page

Type	Prompt
Typo-label-1	How would you describe the emotion you infer from the following text using one of the labels 'prde', 'sadnes', 'bordom', 'anger', 'joy', 'suprise', 'fear', 'guilt', 'disgst', 'turst', 'relif', 'shame', or 'no-emotin'? `{Inst.}`
Typo-label-2	How would you describe the emotion you infer from the following text using one of the labels 'pride', 'sadness', 'boredom', 'angre', 'joy', 'surprise', 'feer', 'guillt', 'disgust', 'trust', 'relief', 'shaem', or 'no-emotion'? `{Inst.}`
Typo-label-3	How would you describe the emotion you infer from the following text using one of the labels 'pried', 'sadnes', 'bordom', 'angr', 'joy', 'suprise', 'feer', 'guillt', 'disgst', 'turst', 'relif', 'shaem', or 'noemotion'? `{Inst.}`
Cap-task	How would you describe the EMOTION you infer from the following TEXT using one of the labels 'pride', 'sadness', 'boredom', 'anger', 'joy', 'surprise', 'fear', 'guilt', 'disgust', 'trust', 'relief', 'shame', or 'no-emotion'? `{Inst.}`
Cap-label	How would you describe the emotion you infer from the following text using one of the labels 'PRIDE', 'SADNESS', 'BOREDOM', 'ANGER', 'JOY', 'SURPRISE', 'FEAR', 'GUILT', 'DISGUST', 'TRUST', 'RELIEF', 'SHAME', or 'NO-EMOTION'? `{Inst.}`
PM-1	How would you describe the emotion you infer from the following text using one of the labels pride, sadness, boredom, anger, joy, surprise, fear, guilt, disgust, trust, relief, shame, or no-emotion? `{Inst.}`
PM-2	How would you describe the emotion you infer from the following text using one of the labels: 'pride', 'sadness', 'boredom', 'anger', 'joy', 'surprise', 'fear', 'guilt', 'disgust', 'trust', 'relief', 'shame', or 'no-emotion'? `{Inst.}`
PM-3	How would you describe the emotion you infer from the following text using one of the labels 'pride'; 'sadness'; 'boredom'; 'anger'; 'joy'; 'surprise'; 'fear'; 'guilt'; 'disgust'; 'trust'; 'relief'; 'shame'; or 'no-emotion'? `{Inst.}`
AltLab-keep-1	How would you describe the emotion you infer from the following text using one of the labels 'self-satisfaction', 'sorrow', 'disinterest', 'rage', 'happiness', 'amazement', 'panic', 'remorse', 'revulsion', 'belief', 'comfort', 'embarrassment', or 'neutral'? `{Inst.}`
AltLab-keep-2	How would you describe the emotion you infer from the following text using one of the labels 'self-respect', 'grief', 'monotony', 'fury', 'delight', 'astonishment', 'terror', 'regret', 'aversion', 'confidence', 'release', 'humiliation', or 'emotionless'? `{Inst.}`
AltLab-int	How would you describe the emotion you infer from the following text using one of the labels 'pride', 'sadness', 'boredom', 'anger', 'joy', 'surprise', 'fear', 'guilt', 'disgust', 'trust', 'relief', or 'shame'? `{Inst.}`
Conf	How would you describe the emotion you infer from the following text using one of the labels 'pride', 'sadness', 'boredom', 'anger', 'joy', 'surprise', 'fear', 'guilt', 'disgust', 'trust', 'relief', 'shame', or 'no-emotion'? Provide the label along with a confidence score (1-5) of your selection, and omit any justification.
Exp	How would you describe the emotion you infer from the following text using one of the labels 'pride', 'sadness', 'boredom', 'anger', 'joy', 'surprise', 'fear', 'guilt', 'disgust', 'trust', 'relief', 'shame', or 'no-emotion'? Provide the label along with a justification for your selection.
Emo-1	How would you describe the emotion you infer from the following text using one of the labels 'pride', 'sadness', 'boredom', 'anger', 'joy', 'surprise', 'fear', 'guilt', 'disgust', 'trust', 'relief', 'shame', or 'no-emotion'? Important: Trust your gut reaction! `{Inst.}`
Emo-2	How would you describe the emotion you infer from the following text using one of the labels 'pride', 'sadness', 'boredom', 'anger', 'joy', 'surprise', 'fear', 'guilt', 'disgust', 'trust', 'relief', 'shame', or 'no-emotion'? Warning: Some texts may be disturbing. `{Inst.}`
Emo-3	How would you describe the emotion you infer from the following text using one of the labels 'pride', 'sadness', 'boredom', 'anger', 'joy', 'surprise', 'fear', 'guilt', 'disgust', 'trust', 'relief', 'shame', or 'no-emotion'? Please be careful with your selection. `{Inst.}`
Emo-4	How would you describe the emotion you infer from the following text using one of the labels 'pride', 'sadness', 'boredom', 'anger', 'joy', 'surprise', 'fear', 'guilt', 'disgust', 'trust', 'relief', 'shame', or 'no-emotion'? Important: Choose quickly and go with your first reaction! `{Inst.}`
Def-task	How would you describe the emotion you infer from the following text using one of the labels 'pride', 'sadness', 'boredom', 'anger', 'joy', 'surprise', 'fear', 'guilt', 'disgust', 'trust', 'relief', 'shame', or 'no-emotion'? Your choice should reflect one primary emotion you infer the author or speaker is expressing, based on the tone, context, and implied sentiment of the text. `{Inst.}`

Continued on next page

Table 8 – continued from previous page

Type	Prompt
Def-label-1	How would you describe the emotion you infer from the following text using one of these labels? 'pride': A sense of self-respect or accomplishment. 'sadness': Emotional pain, sorrow, or unhappiness. 'boredom': Lack of interest or engagement. 'anger': Strong displeasure or hostility. 'joy': Intense happiness or delight. 'surprise': Unexpectedness or astonishment. 'fear': Anxiety or distress caused by threat. 'guilt': Remorse over wrongdoing. 'disgust': Revulsion or strong disapproval. 'trust': Confidence in reliability or honesty. 'relief': Reassurance after distress. 'shame': Humiliation or embarrassment. 'no-emotion': Neutral tone; absence of discernible feeling. `{Inst.}`
Def-label-2	How would you describe the emotion you infer from the following text using one of these labels? 'pride': A feeling of self-respect, achievement, or personal worth. 'sadness': A state of unhappiness, grief, or emotional pain. 'boredom': A lack of interest or engagement; feeling unstimulated. 'anger': A strong feeling of displeasure or hostility. 'joy': A feeling of great happiness or delight. 'surprise': A reaction to something unexpected or sudden. 'fear': An emotional response to threat or danger, real or perceived. 'guilt': A feeling of responsibility or remorse for a wrongdoing. 'disgust': A sense of revulsion or profound disapproval. 'trust': Confidence or belief in the reliability or integrity of someone/something. 'relief': A feeling of reassurance or alleviation of distress. 'shame': A painful feeling of humiliation or embarrassment due to perceived wrongdoing. 'no-emotion': Used when the text is emotionally neutral or lacks a discernible emotional tone. `{Inst.}`
Def-both	How would you describe the emotion you infer from the following text using one of these labels? Your choice should reflect one primary emotion you infer the author or speaker is expressing, based on the tone, context, and implied sentiment of the text. 'pride': A feeling of self-respect, achievement, or personal worth. 'sadness': A state of unhappiness, grief, or emotional pain. 'boredom': A lack of interest or engagement; feeling unstimulated. 'anger': A strong feeling of displeasure or hostility. 'joy': A feeling of great happiness or delight. 'surprise': A reaction to something unexpected or sudden. 'fear': An emotional response to threat or danger, real or perceived. 'guilt': A feeling of responsibility or remorse for a wrongdoing. 'disgust': A sense of revulsion or profound disapproval. 'trust': Confidence or belief in the reliability or integrity of someone/something. 'relief': A feeling of reassurance or alleviation of distress. 'shame': A painful feeling of humiliation or embarrassment due to perceived wrongdoing. 'no-emotion': Used when the text is emotionally neutral or lacks a discernible emotional tone. `{Inst.}`



Welcome to the Study: Offensiveness in Social Media

Dear participant,

Thank you for your interest in our study. This study investigates how people perceive offensiveness in language.

In this study, we will ask you to annotate 22 social media comments. We will also collect some demographic information, including your age, gender, occupation status, and field of occupation.

Task Details

The study should take you about 8 minutes to complete. You will receive a reward of 1.25 £ for your participation. Participation is voluntary.

To take part, you must be at least 18 years old and a native speaker of English. You may withdraw at any time without providing a reason. However, please note that you will not be compensated if you choose to withdraw. The use of AI tools is not allowed. You will only receive payment if you complete the study fully, your responses are meaningful, and you pass all attention checks.

Privacy

All data collected will be used for research purposes only and will be anonymized before being shared publicly. We may include anonymized examples from the data in scientific publications.

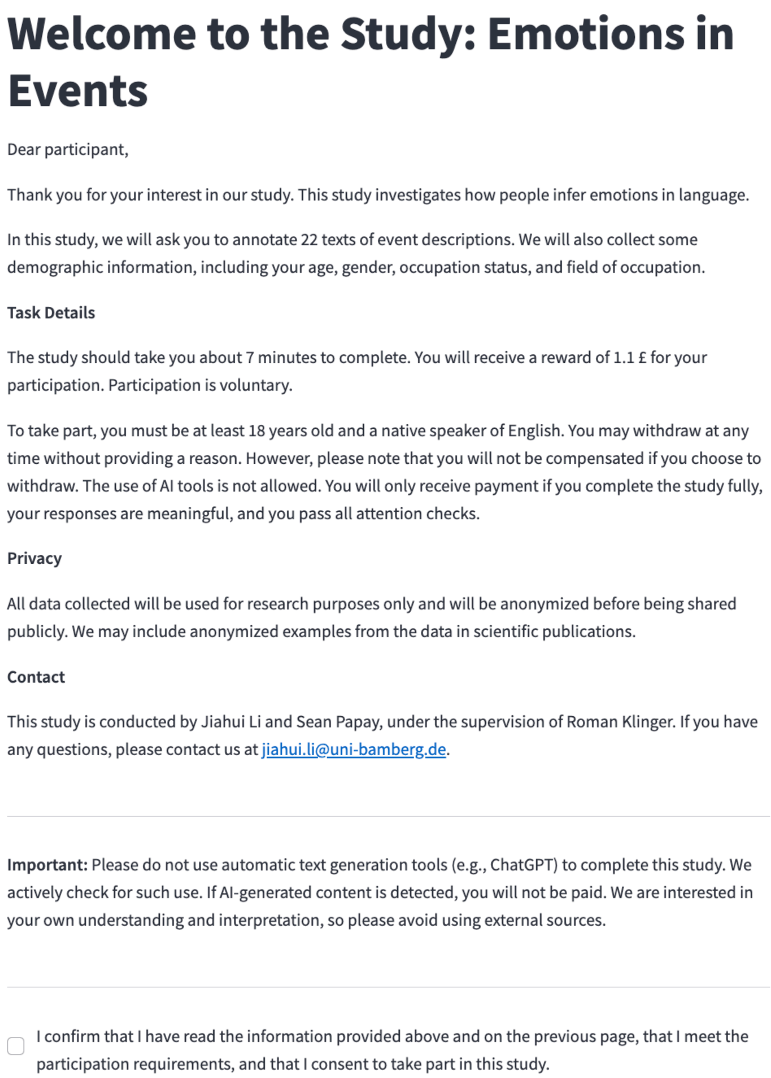
Contact

This study is conducted by Jiahui Li and Sean Papay, under the supervision of Roman Klinger. If you have any questions, please contact us at jiahui.li@uni-bamberg.de.

Important: Please do not use automatic text generation tools (e.g., ChatGPT) to complete this study. We actively check for such use. If AI-generated content is detected, you will not be paid. We are interested in your own understanding and interpretation, so please avoid using external sources.

☐ I understand that some of the content I may encounter could be offensive or upsetting. I confirm that I have read the information provided above and on the previous page, that I meet the participation requirements, and that I consent to take part in this study.

Figure 5: A screenshot of the instruction page for offensiveness rating survey.



Welcome to the Study: Emotions in Events

Dear participant,

Thank you for your interest in our study. This study investigates how people infer emotions in language.

In this study, we will ask you to annotate 22 texts of event descriptions. We will also collect some demographic information, including your age, gender, occupation status, and field of occupation.

Task Details

The study should take you about 7 minutes to complete. You will receive a reward of 1.1 £ for your participation. Participation is voluntary.

To take part, you must be at least 18 years old and a native speaker of English. You may withdraw at any time without providing a reason. However, please note that you will not be compensated if you choose to withdraw. The use of AI tools is not allowed. You will only receive payment if you complete the study fully, your responses are meaningful, and you pass all attention checks.

Privacy

All data collected will be used for research purposes only and will be anonymized before being shared publicly. We may include anonymized examples from the data in scientific publications.

Contact

This study is conducted by Jiahui Li and Sean Papay, under the supervision of Roman Klinger. If you have any questions, please contact us at jiahui.li@uni-bamberg.de.

Important: Please do not use automatic text generation tools (e.g., ChatGPT) to complete this study. We actively check for such use. If AI-generated content is detected, you will not be paid. We are interested in your own understanding and interpretation, so please avoid using external sources.

☐ I confirm that I have read the information provided above and on the previous page, that I meet the participation requirements, and that I consent to take part in this study.

Figure 6: A screenshot of the instruction page for the emotion classification survey.

Figure 7: A screenshot example of the annotation task for the emotion classification survey. The prompt variant is instantiated with a typological error modification (Typo-label-3).