The Overlooked Repetitive Lengthening Form in Sentiment Analysis

Lei Wang and **Eduard Dragut**
Temple University, Philadelphia, PA, USA
{tom.lei.wang,edragut}@temple.edu

Abstract

Individuals engaging in online communication frequently express personal opinions with informal styles (e.g., memes and emojis). While Language Models (LMs) with informal communications have been widely discussed, a unique and emphatic style, the Repetitive Lengthening Form (RLF), has been overlooked for years. In this paper, we explore answers to two research questions: 1) Is RLF important for sentiment analysis (SA)? 2) Can LMs understand RLF? Inspired by previous linguistic research, we curate **Lengthening**, the first multi-domain dataset with 850k samples focused on RLF for SA. Moreover, we introduce **Exp**lainable **Instruct**ion Tuning (**ExpInstruct**), a two-stage instruction tuning framework aimed to improve both performance and explainability of LLMs for RLF. We further propose a novel unified approach to quantify LMs' understanding of informal expressions. We show that RLF sentences are expressive expressions and can serve as signatures of document-level sentiment. Additionally, RLF has potential value for online content analysis. Our results show that fine-tuned Pre-trained Language Models (PLMs) can surpass zero-shot GPT-4 in performance but not in explanation for RLF. Finally, we show ExpInstruct can improve the open-sourced LLMs to match zero-shot GPT-4 in performance and explainability for RLF with limited samples. Code and sample data are available at https://github.com/Tom-Owl/OverlookedRLF

1 Introduction

Informal styles are prevalent on social media platforms, where people use nuanced expressions to share opinions and emotions personally and engagingly (Yang et al., 2020; Hosseinia et al., 2021; He et al., 2019, 2021). Previous research has explored various informal styles such as meme (Lin et al., 2024; Sharma et al., 2023), emoji (Peng et al., 2023; Barbieri et al., 2018; Reelfs et al., 2022), slogan (Iwama and Kano, 2018; Misawa et al., 2020) and abbreviation (Gorman et al., 2021). However, it remains a challenge for Language Models (LMs) to understand the sentiment in nuanced and subtle linguistic expressions, which require a deep contextual and cultural understanding (Zhang et al., 2023b). This work delves into one specific informal expression - Repeated Lengthening Form (RLF), which refers to the linguistic phenomenon where additional characters are added to the standard spelling of a word to enhance or alter its conveyed meaning (Brody and Diakopoulos, 2011; Kalman and Gergle, 2014). We further generalize the concept of RLF and divide it into two types: Repetitive Letters (e.g., 'loooove') and Repetitive Punctuations (e.g., 'love!!!!'). Our study finds that an average of 5.8% documents possess RLF among 4 public datasets and 5 domains (Table 1) where some of them include more than 13% documents with RLF.

LMs consist of Pre-trained Language Models (PLMs) (Zhao et al., 2023) such as RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2019), and GPT-2 (Radford et al., 2019), and Large Language Models (LLMs), including GPT-4 (OpenAI, 2023) and LLaMA2 (Touvron et al., 2023). LMs exhibit impressive performance on many NLP tasks (Brown et al., 2020; Wang et al., 2023). However, we know little about the boundaries of performance and explainability of LMs for RLF. Such linguistic features, common in daily communication, have yet to be thoroughly investigated (Go et al., 2009; Nguyen and Nguyen, 2017; Abdul-Mageed and Ungar, 2017; Ali et al., 2019; Aljebreen et al., 2021). The lack of a specialized dataset for RLF impedes us from evaluating and improving LMs to learn the nuanced communications in real-world and online social media content. This research gap raises two research questions: **1) Is RLF important for SA? 2) Can LMs understand RLF?** In this study, we aim to evaluate and improve the performance and explainability of PLMs and LLMs for RLF.

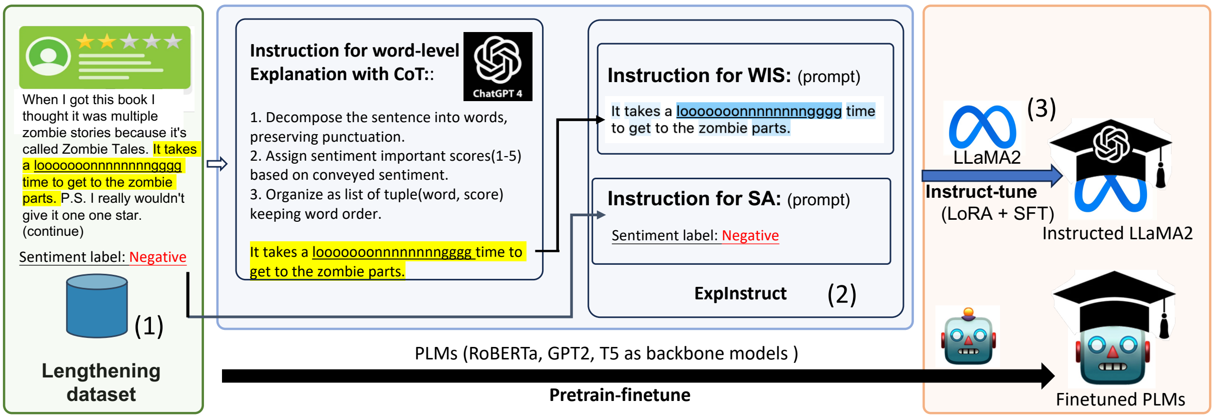

When I got this book I thought it was multiple zombie stories because it's called Zombie Tales. It takes a looooooonnnnnnnnngggg time to get to the zombie parts. P.S. I really wouldn't give it one one star. (continue)
Sentiment label: Negative
(1)
Lengthening dataset
Instruction for word-level Explanation with CoT::
ChatGPT 4
1. Decompose the sentence into words, preserving punctuation.
2. Assign sentiment important scores(1-5) based on conveyed sentiment.
3. Organize as list of tuple(word, score) keeping word order.
It takes a looooooonnnnnnnnngggg time to get to the zombie parts.
Instruction for WIS: (prompt)
It takes a looooooonnnnnnnnngggg time to get to the zombie parts.
Instruction for SA: (prompt)
Sentiment label: Negative
ExpInstruct
(2)
(3)
LLaMA2
Instruct-tune
(LoRA + SFT)
Instructed LLaMA2
PLMs (RoBERTa, GPT2, T5 as backbone models)
Pretrain-finetune
Finetuned PLMs

Figure 1: An overview of our work for RLF. (1) We introduce **Lengthening** in Section 3. (2) We propose the **ExpInstruct** framework and describe prompt details in Section 4.1. (3) Experiments details are in Section 5.

The overview of our work is shown in Figure1. To answer the first research question, we curate **Lengthening**, the first dataset focused on RLF for SA. Inspired by previous linguistic research of RLF (Kalman and Gergle, 2014; Gray et al., 2020), we design a pipeline for extracting RLF sentences and words from 4 public datasets and sample 850k instances. We conduct comprehensive experiments to compare zero-shot performance with 3 PLMs and 2 LLMs between RLF and w/o RLF groups. Our results reveal the sentiment-expressive value of RLF sentences with consistent higher performance. Importantly, we demonstrate the transferability of our fine-tuned LMs with **Lengthening** get document-level gain. Specifically, we observe an average improvement of a 7.6% in accuracy and 4.5% in F1 score (Table 8). These observations collectively show that RLF sentences can serve as key sentences for document-level SA tasks. In addition, we highlight the potential of RLF for online social media content analysis, where short text and informal communications prevail.

We study the second question and show that PLMs can reach better performances than zero-shot GPT-4 after fine-tuning on **Lengthening**, but lag behnd in explainability. We solve this issue with **ExpInstruct**, which can empower LLMs to reach the same level of performance and explainability as zero-shot GPT-4 with a small-size subset dataset. We analyze the data quality and explanation reliability with human evaluation. We further explore the effect of sample size and instruction strategy for ExpInstruct with ablation study.

In summary, our contributions are as follows:

- We call attention to RLF, an overlooked linguistic informal style. We show that RLF sentences can serve as signatures of document sentiment and have potential value for user generated content analysis.
- We introduce **Lengthening**, a multi-domains dataset featuring RLFs with 850k samples grounding from 4 public datasets for SA tasks.
- We propose a cost-effective approach **ExpInstruct**, which can improve the performance and explainability of open-sourced LLMs for RLF to the same level of zero-shot GPT-4.
- We quantify the explainability of PLMs and LLMs for RLF with a unified approach. Human evaluation demonstrates the reliability of this method.

2 Related Work

Repetitive Lengthening Form (RLF) The study by (Kalman and Gergle, 2014) shows that RLF is a written emulation of nonverbal spoken cues in user-generated content and proposes a method to reduce stretched words to their root words. Brody and Diakopoulos (2011) highlights the importance of accurately interpreting RLF for SA and employing RLF to augment sentiment dictionary (Dragut et al., 2010). Schnoebelen et al. (2012); Gray et al. (2020) describe this nuanced linguistic feature as expressive words often used to emphasize or exaggerate the underlying sentiment intensity of the root word. Previous research on RLF has primarily focused on linguistic and statistical analysis. A systematic study of the sentiment value of RLF in transform-based LMs is still lacking. We curate

Domain	Dataset	Original Tag	#Document	RLF Ratio(%)	#Samples	Label Distribution (1/0)	#Unique RLF Word	#Unique Root Word
Books	Amazon Review	1-5 stars	51,312k	5.41	255k	92/8(%)	11,813	1,592
Electronics	Amazon Review	1-5 stars	20,994k	4.32	212k	74/26(%)	11,367	1,541
Restaurants	Yelp	1-5 stars	6,990k	11.28	315k	74/26(%)	11,642	1,587
Social Media	Twitter	binary	1,600k	13.36	44k	46/54(%)	5,337	1,149
Hotels	TripAdvisor	1-5 stars	879k	10.01	24k	76/24(%)	4,558	1,250
ALL	**Lengthening**	binary	850k	100	850k	78/22(%)	19,610	1,677

Table 1: Summary statistics for our dataset. **Lengthening** is the first large-scale dataset featuring RLF for SA task, grounded in 4 public datasets with an average of 5.8% documents containing RLF. More details of the generation process of our dataset are described in Section 3

Domain	Lengthening Style	POS	Lengthening Word	Root Word	Label	Example Sentence
Books	Punctuation	Noun	book!!!!!	book!	1	Do yourself a favour a read this book!!!!!
Electronics	Letter	Verb	loooove	love	1	I loooove my new phone case.
Restaurants	Punctuation	Adj	amazing!!!!!	amazing!	1	We are from Seattle and this coffee is amazing!!!!!
Twitter	Letter	Adv	SOOOO	SO	0	SOOOO bummed i'm going to miss sam's party tonight.
Hotel	Punctuation	Noun	year............	year...	0	I am looking to go back next year............

Table 2: Samples from our **Lengthening** dataset.

a large-scale dataset **Lengthening** with RLF for SA. We use the dataset to evaluate and improve the performance and explainability of the up-to-date PLMs and LLMs for RLF.

Instruction Tuning is a paradigm that fine-tunes language models on multiple tasks with instruction-input-output pairs to improve performance and generalize to unseen tasks (Wei et al., 2021; Sanh et al., 2022). Emerging work explores instruction tuning to tasks such as text editing (Raheja et al., 2023), information extraction (Lu et al., 2023) and classification (Aly et al., 2023). We extend this line of work by instruct-tuning LLMs for SA with RLF. We focus on a novel task of predicting the document-level sentiment label using only one sentence.

Honovich et al. (2023); Yin et al. (2023) use in-context learning strategies to prompt LLMs to automatically generate data for instruction tuning. Lampinen et al. (2022); Zhang et al. (2023c); Zhou et al. (2023) instruct LLMs with emphasis on explainability. Inspired by those works, we prompt GPT-4 with Chain of Thought (CoT) (Wei et al., 2023) to generate word importance scores for RLF sentences. Using these scores along with the ground truth document-level sentiment labels, we automatically generate prompts for explainable instruction tuning of LLMs.

Model Explanation The importance of input features has been extensively explored in recent years. Studies such as (Ribeiro et al., 2016; Li et al., 2017; Zhu et al., 2024; Ray Choudhury et al., 2022; Atanasova et al., 2020) have employed saliency-based methods for PLMs which are based on changes in loss or gradient metrics. The approach does not apply to prompt-interaction LLMs. Furthermore, the loss or gradient values are not always available for closed-source LLMs (e.g., GPT-4 and Claude). Prompting methods with CoT are proposed for feature importance analysis with LLMs (Zhong et al., 2023; Wang et al., 2023). However, the difference between saliency and prompt based methods creates a barrier to comparing explainability results between PLMs and LLMs. We overcome that challenge by proposing a unified approach for evaluating the explainability of LMs in handling RLF (Section 4.2).

3 The Lengthening Dataset

We introduce the **Lengthening** dataset in this section. We need a few definitions first: RLF sentence is a sentence with one or more RLF words, RLF document is a document (e.g., user review) with at least one RLF sentence. We present an overview of data statistics for our dataset and present in Table1. More detailed examples from our dataset can be found in Table 2.

3.1 Data Source

To comprehensively evaluate the usage of RLF in social media platforms and online user reviews, we select 4 public datasets covering 5 distinct domains: **Books & Electronics** from Amazon Reviews (Ni et al., 2019); **Restaurant Reviews** from Yelp (Yelp, 2021) data from Feb 16, 2021; **Hotel Reviews** from TripAdvisor (Li, 2020); **Twitter** dataset with general social posts (Go et al., 2009). All user reviews (documents) are categorized based on their star ratings: 1-2 stars as negative, 4-5 stars as positive, with 3-star reviews being excluded as they are assumed to be neutral.

3.2 Generation of Lengthening

We describe our pipeline for extracting RLF sentences and words from documents with three steps. We give additional details about dataset generation techniques and algorithms in Appendix A.

Identification of Potential RLF Documents We design a regular expression, termed RLFsearch ('([a-zA-Z])\1{2,}|[!]{3,}|[?]{3,}|[,]{3,}|[.]{4,}') to identify potential RLF documents (Gray et al., 2020). Documents that return a positive result from RLFsearch are retained for the next step.

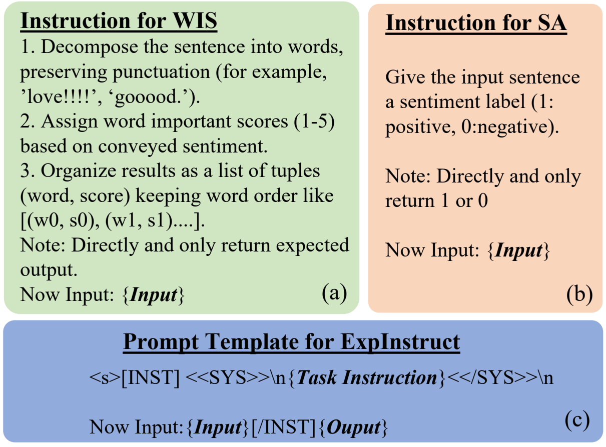
Potential RLF Sentences Extraction We segment potential RLF documents into sentences. Sentences containing fewer than five words are merged with the preceding sentences to maintain the overall sentiment polarity (Dragut et al., 2012, 2015). Each sentence is re-evaluated using the RLFsearch regex to extract potential RLF sentences.

RLF Extraction We further split potential RLF sentences into individual words and apply the RLFsearch regex at word-level. This step excludes numbers, URLs, words beginning with '@', and monetary amounts. Words that pass this regex filter are recognized as RLF words. Sentences that contain one or more RLF words are extracted as RLF sentences. Correspondingly, documents with one or more RLF sentences are extracted as RLF documents. For each RLF sentence, a sentence without any RLFs (w/o RLF) from the same document is randomly chosen for zero-shot comparison (we only consider documents containing two or more sentences). In addition, we find the root word of each RLF word with the algorithm proposed in (Kalman and Gergle, 2014) based on American English. For example, RLF words like 'loooove' and 'loooovvve' have the same root word 'love'.

4 Method

We introduce ExpInstruct, a two-stage instruction tuning framework aimed to improve both performance and explainability of LLMs for RLF. ExpInstruct first prompts GPT-4 for WIS scores and then finetunes LLMs with instructions for WIS and SA. We further propose a unified approach to evaluate the comprehension level of LMs for RLF.

Formally, consider (x, rlf, y) to be a single tuple in our dataset D, where x is an RLF sentence with an RLF word rlf and y is the document-level sentiment label. We denote a transformer-based model by f.



Instruction for WIS

1. Decompose the sentence into words, preserving punctuation (for example, 'love!!!!', 'gooood.').
2. Assign word important scores (1-5) based on conveyed sentiment.
3. Organize results as a list of tuples (word, score) keeping word order like [(w0, s0), (w1, s1)....].

Note: Directly and only return expected output.
Now Input: {***Input***}

(a)

Instruction for SA

Give the input sentence a sentiment label (1: positive, 0:negative).

Note: Directly and only return 1 or 0

Now Input: {***Input***}

(b)

Prompt Template for ExpInstruct

<s>[INST] <<SYS>>\n{***Task Instruction***}<</SYS>>\n

Now Input:{***Input***}[/INST]{***Ouput***}

(c)

Figure 2: Prompt Design and Template for ExpInstruct. (a) Prompt with CoT for word-level explainability. (b) Simple Prompt for SA. (c) Prompt Template for Instruction tuning

4.1 ExpInstruct

Prompt Design for Explanation We prompt GPT-4 with CoT to generate Word Importance Scores (WIS) to reflect word-level understanding of input sentence x as shown in Figure 2(a). The CoT consists of 3 sequential reasoning steps. 1) Sentence Decomposition: Segment sentences into words and keep punctuation marks with a few-shot strategy. 2) Word Importance Scoring: Assign sentiment importance scores (1-5) to each word, which can reflect LLMs' understanding of word-level sentiment. 3) Structured Output: We specify the structured output format for subsequent analysis.

Instruction Template ExpInstruct has two tasks with the same prompt template as shown in Figure 2(c). This is achieved by adding three placeholders: {Task Instruction}, {Input}, and {Output}. The first task is *Instruction for WIS*: The Task Instruction is shown in Figure 2(a), with the Input as RLF sentence x and the Output is the structured output generated by GPT-4 for WIS. The second task is *Instruction for SA*: The Task Instruction is shown in Figure 2(b); the Input is the RLF sentence x and the Output is the document-level sentiment label y.

4.2 A Unified Approach to Evaluate Explainability

In this section, we propose a unified approach to evaluate the explainability of PLMs and LLMs with the help of WIS. Our approach consists of two steps: 1) Generate WIS from LLMs and PLMs, and 2) Quantify explainability across models with normalization.

Generate WIS For LLMs, we use a prompt-based method to generate WIS (Zhong et al., 2023; Wang et al., 2023) with the instruction shown in Figure 2(a). This method requires only one-time inference for each input sentence x and is label-free:

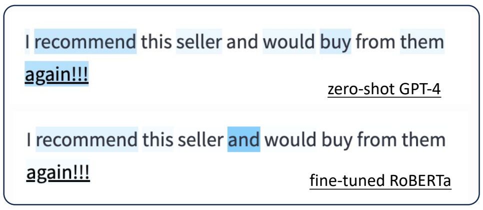
$$\text{WIS} = f(x, \text{Instruction for WIS}) \quad (1)$$

For PLMs, we choose a saliency-based method to generate WIS. Specifically, we use the occlusion-based method (Ray Choudhury et al., 2022; Zhu et al., 2024) because it's intuitive and applicable to various PLMs. This method involves sequentially removing one word w_i to observe the absolute change in the loss value, serving as an indicator of the word-level significance:

$$\text{WIS}[w_i] = |L[f(x), y] - L[f(x - w_i), y]| \quad (2)$$

where $f(x)$ represents the output logit value from the transformer-based model f given the input x, L is the cross-entropy loss function, and $x - w_i$ is the sentence after the removal of the word w_i from the input sentence.

Quantify Explainability To eliminate the barrier caused by differing relative WIS values across models, we apply min-max followed by L_1 normalization to WIS and denote the normalized WIS as WIS_{norm}. We visualize the normalized WIS from zero-shot GPT-4 and fine-tuned RoBERTa with a sample sentence in Figure 3.



I recommend this seller and would buy from them again!!!

zero-shot GPT-4

I recommend this seller and would buy from them again!!!

fine-tuned RoBERTa

Figure 3: Comparing normalized WIS for an RLF sentence from zero-shot GPT-4 and fine-tuned RoBERTa.

The overall explainability score S_{exp} for a given dataset D can be quantified as:

$$S_{exp} = \frac{1}{|D|} \sum_{(x,rlf,y)\in D} \text{WIS}_{\text{norm}} \left[j_{(w_j = rlf)} \right] \quad (3)$$

where j refers to the word index where $w_j = rlf$ in the input sentence x.

We take an average of the WIS_{norm} of RLF for all (x, rlf, y) pairs to compute the explainability score S_{exp} for the target model f on dataset D. A higher S_{exp} score indicates that the model pays more attention to RLF words for the SA task, which reflects a better understanding of the expressive value of RLF. This metric enables us to automatically quantify the explainability of LMs for RLF. It can be used as a complement to the qualitative method which relies on case studies and human verification (Ray Choudhury et al., 2022; Zhu et al., 2024).

5 Experimental Setup

In this section, we introduce baseline models, experimental design and implementation details.

Baseline Models To explores the boundaries of performance and explainability for SOTA PLMs and LLMs for RLF, we choose 3 fine-tuned 3 PLMs for SA task with backbone models as RoBERTa (Large), GPT-2 (Medium) and T5 (Base). These models have comparable parameter scales and represent encoder-only, decoder-only, and encoder-decoder architectures. For LLMs, we use GPT-4 via the OpenAI API. Moreover, we select LLaMA2 (13B-chat-hf) for instruct-tuning because it's open-sourced with auto-regressive architecture like GPT-4. And the scalable parameter size of LLaMA2 (13B) allows inference and fine-tuning with LoRA (Hu et al., 2021) on a single GPU. More details of these baseline models and API usage are provided in the Appendix B.1

Implementation We fine-tune 3 PLMs on the entire Lengthening dataset. The batch sizes are set to 64 for RoBERTa and T5, and 32 for GPT-2 to optimize GPU memory usage. We sampled 3,000 instances from Lengthening as a subset dataset for experiments with LLMs with stratified sampling method based on domains, where we used 1.6k instances for training and others for validation and testing. This sample size aligns with previous studies (Wang et al., 2023; Zhang et al., 2023a; Deng et al., 2022). More details of data split see Appendix A.2.

We instruct-tune LLaMA2 with the our ExpInstruct framework with LoRA and Supervised fine-tuning (SFT) (von Werra et al., 2020) to lower computational costs. We set the temperature at 0.2 for all LLMs. All models are trained for a maximum of 5 epochs with k-fold cross-validation strategy (k = 3) to identify the best checkpoints. In each iteration one fold serves as the test set while the remaining data is split into train/val with 4/1 ratio. It takes one week to conduct all experiments with a single RTX A6000 GPU with 50 GB of memory.

Metric	Acc(%)		F1(%)	
Backbone Model	RLF	w/o RLF	RLF	w/o RLF
RoBERTa (Large)	85.94 ± 0.03	84.40 ± 0.04	**90.67** ± 0.04	89.56 ± 0.03
GPT-2 (Medium)	79.56 ± 0.07	77.56 ± 0.08	85.76 ± 0.08	84.22 ± 0.06
T5 (Base)	83.22 ± 0.06	81.93 ± 0.04	88.71 ± 0.07	87.85 ± 0.03
LLaMA2 (13B)	76.25 ± 1.76	70.41 ± 0.57	84.30 ± 1.28	82.64 ± 0.39
GPT-4	**86.26** ± 0.93	**86.20** ± 1.45	90.12 ± 0.75	**90.08** ± 1.14

Table 3: Overall zero-shot accuracy and F1 score for sentences with RLF and without RLF words (w/o RLF). Bold denotes the best results in a column and underline highlights the second best results. ± indicates standard deviation score.

6 Results

Is RLF Important for Sentiment Analysis? We approach this question from a new angle: predicting the document-level sentiment label with a single sentence. Specifically, we compare the performance of two groups: sentences with RLF and those without (w/o RLF). Evaluations are conducted for zero-shot and fine-tuned models with Accuracy (Acc) and macro F1 score (F1) as performance metrics.

Firstly, we present the zero-shot results in Table 3. The RLF group consistently achieves better performance than the w/o RLF group in both Accuracy and F1 score across all models. These results indicate that sentences with RLF can serve as key signatures for document-level sentiment. To our knowledge, this study is the first to empirically demonstrate the sentiment-expressive value of RLF through comprehensive experiments with various PLMs and LLMs.

Furthermore, we conduct a domain-wise analysis of zero-shot performance as shown in Table 4. Due to sample size limitations for the subset dataset, this analysis is focused on PLMs. The superiority of the RLF sentences in expressing sentiment is consistently observed across diverse domains and models, highlighting the robustness and generalizability of the sentiment-expressive value of RLF.

In Figure 4, we present the accuracy of zero-shot and fine-tuned PLMs in relation to sentence length. We observe performance improvements in all PLMs for both the RLF and w/o RLF groups after fine-tuning on our **Lengthening** dataset. Superisingly, the RLF group demonstrates significantly better performance than the w/o RLF group when sentence lengths are within 80 characters, with more than 70% of sentences within this range. This performance gap is evident in both zero-shot and fine-tuned models. These findings highlight the critical importance of focusing on RLF in social media content, where short and informal expressions prevail.

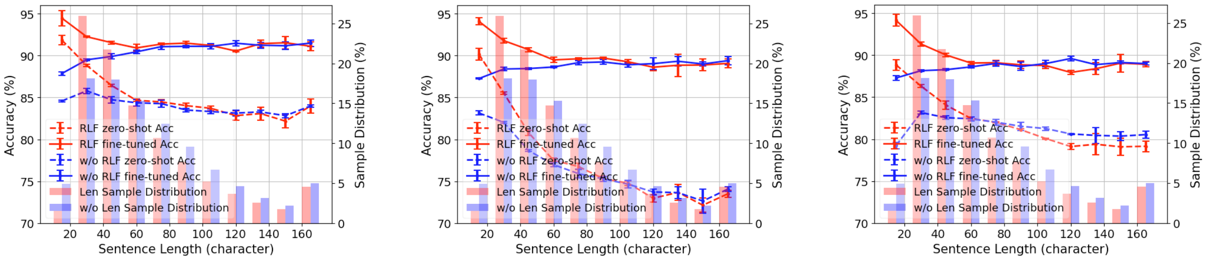
Can LMs Understand RLF? To answer this question, we compare the performance and explainability of zero-shot and fine-tuned models for RLF sentences with Acc, F1 and S_{exp} as evaluation metrics. We present the results for two RLF styles in Table5. Our results show that GPT-4 has the best performance and explainability of RLF among zero-shot models. Although zero-shot GPT-2 (Medium) has the highest S_{exp} score, its low Acc and F1 scores suggest an insufficient understanding of RLF. This is further supported by the drop of S_{exp} score after fine-tuning GPT-2 (Medium). We observe that all fine-tuned PLMs achieve better Acc and F1 scores compared to zero-shot GPT-4. However, their S_{exp} scores are still lower than zero-shot GPT-4 with a significant gap, suggesting that the fine-tuned PLMs may lack sufficient understanding of RLF. This appears to be another instance of LMs being "right for the wrong reasons"(McCoy et al., 2019).

Interestingly, our ExpInstruct with LLaMA2 achieves the same level of performance and explainability as zero-shot GPT-4 with only 1,600 instruction samples. This finding highlights the effectiveness of ExpInstruct, a cost-effective approach that requires only limited samples for instruction tuning to enhance open-source LLMs as alternatives to GPT-4.

7 Analysis

In this section, we first conduct human evaluation for data quality and explanation reliability. Additionally, we verify the benefits gained from **Lengthening** are transferable to document-level SA. Moreover, ablation studies are conducted to explore the effects of training sample size and instruct strategy.

Data Quality The final sentiment label for each sentence was determined by majority vote among 3 annotators. We use Krippendorff's Alpha score (Krippendorff, 2018) for inter-rater agreement (IAA) score and obtained a score of 0.86, indicating that the annotated data is reliable. We report human performance on our sentence to document SA task with Acc / F1 score as 90.01 / 92.52% for the RLF group and 85.50 / 89.04% for the w/o RLF group. The superior scores of the RLF group support our main conclusion that RLF sentences can serve as



(a) RoBERTa (Large)

(b) GPT-2 (Medium)

(c) T5 (Base)

Figure 4: Comparison of accuracy between the RLF and w/o RLF groups using zero-shot and fine-tuned models by sentence length. The lines represent average values, and the error bar indicate the standard deviation for each length group across 3 runs. Both results across 3 fine-tuned models show a convergence between the RLF and w/o RLF groups when the sentence character length is around 80.

Backbone Model	RoBERTa (Large)				GPT-2 (Medium)				T5 (Base)			
Metric	Acc(%)		F1(%)		Acc(%)		F1(%)		Acc(%)		F1(%)	
Domain / Group	RLF	w/o RLF	RLF	w/o RLF	RLF	w/o RLF	RLF	w/o RLF	RLF	w/o RLF	RLF	w/o RLF
Books	**87.36** ± 0.17	86.90 ± 0.16	**92.76** ± 0.10	92.46 ± 0.10	80.65 ± 0.07	81.21 ± 0.18	88.45 ± 0.06	88.84 ± 0.11	86.09 ± 0.17	87.07 ± 0.08	92.01 ± 0.10	92.63 ± 0.05
Restaurants	**88.46** ± 0.07	85.93 ± 0.09	**92.07** ± 0.07	90.24 ± 0.06	82.67 ± 0.20	79.04 ± 0.02	87.60 ± 0.18	84.70 ± 0.02	85.26 ± 0.23	82.79 ± 0.14	89.70 ± 0.19	87.95 ± 0.09
Electronics	**84.56** ± 0.08	82.54 ± 0.19	**88.86** ± 0.10	87.27 ± 0.16	76.35 ± 0.09	73.28 ± 0.17	81.72 ± 0.16	78.99 ± 0.21	80.53 ± 0.12	77.84 ± 0.05	85.65 ± 0.02	83.54 ± 0.09
Twitter	66.30 ± 0.36	**66.87** ± 0.40	**68.64** ± 0.35	68.41 ± 0.43	66.84 ± 0.13	65.71 ± 0.57	66.11 ± 0.14	65.53 ± 0.98	65.55 ± 0.35	64.56 ± 0.05	66.36 ± 0.50	65.22 ± 0.19
Hotel	84.07 ± 0.18	**84.37** ± 0.60	89.11 ± 0.16	**89.32** ± 0.43	76.55 ± 0.20	76.27 ± 0.49	83.14 ± 0.06	82.97 ± 0.39	80.19 ± 0.30	81.43 ± 0.10	86.19 ± 0.23	87.24 ± 0.02

Table 4: Zero-shot accuracy and F1 score for sentences with/without RLF words in different domains. We report RoBERTa (Large), GPT-2 (Medium) and T5 (Base) because limited test samples for GPT-4 and LLaMA2-13B. Bold and underline indicate best and second best results. ± indicates standard deviation score.

key sentence for document-level SA. This result shows that zero-shot GPT-4 is close to human-level performance in our task. We further present the confusion matrix for sample distribution in Table 6.

Explanation Reliability We ask the annotators to evaluate the reliability of the WIS generated by the zero-shot GPT-4 and the four fine-tuned LMs (RoBERTa, GPT-2, T5, and LLaMA2) with criteria of 1: Agree, 0: Disagree. The final reliability score for each sample was determined by averaging scores among annotators. We report a moderate overall IAA score of 0.44 and show detailed result in Table 7. This result supports our conclusion that fine-tuned PLMs still have a gap in understanding RLF compared to zero-shot GPT-4. Furthermore, the correlation coefficient between the reliability score and S_{exp} is 0.91, showing the validity of our unified approach for explainability evaluation proposed in Section 4.2.

7.1 Human Evaluation

We conduct human evaluation to assess potential errors in our methodology. Specifically, we randomly selected 200 samples and recruited 3 annotators for sentence sentiment label annotation and WIS reliability scores evaluation. More details about human evaluation in Appendix C

7.2 Transferability to Document-level SA

In this section, we explore whether the benefits gained from fine-tuning models on RLF sentences are transferable to document-level SA. Table 8 presents the document-level performance with fine-tuned LMs on the subset dataset (Section 5). The results show that fine-tuned models consistently outperform zero-shot models on document-level SA both in accuracy and F1 scores. The non-overlapping confidence intervals indicate that these improvements are statistically significant. This experiment and Table 5 verify that our **Lengthening** dataset enables LMs to understand better RLF with generalization ability and improvements at the word, sentence, and document levels.

7.3 Ablation Study

Effect of Sample Size We compare the performance of ExpInsturct with different training sample sizes. As shown in Table 9, the best result is achieved with 1600 samples for instruction tuning, which is the data split strategy we chose for the main experiment. We observe a continuous increase in all metrics as the sample size increases, yet the rate of gain slows after the sample size reaches 1000.

Effect of Instruction Strategy We explore how our proposed ExpInsturct strategy helps LLMs im-

Lengthening Style	Punctuation Repetitive			Letter Repetitive			Overall		
Backbone Model	S_{exp}	Acc(%)	F1(%)	S_{exp}	Acc(%)	F1(%)	S_{exp}	Acc(%)	F1(%)
Zero-shot									
RoBERTa (Large)	0.24 ± .001	86.65 ± 0.01	91.19 ± 0.01	0.10 ± .001	**83.61** ± 0.16	**88.93** ± 0.17	0.21 ± .001	85.94 ± 0.03	**90.67** ± 0.04
GPT-2 (Medium)	**0.49** ± .001	80.55 ± 0.11	86.57 ± 0.08	0.07 ± .001	76.34 ± 0.11	82.98 ± 0.15	**0.39** ± .002	79.56 ± 0.07	85.76 ± 0.08
T5 (Base)	0.20 ± .001	83.97 ± 0.04	89.30 ± 0.04	0.11 ± .001	80.75 ± 0.10	86.74 ± 0.16	0.18 ± .001	83.22 ± 0.06	88.71 ± 0.07
LLaMA2 (13B)	0.18 ± .004	77.25 ± 2.62	85.23 ± 1.89	0.25 ± .003	73.54 ± 3.86	81.46 ± 2.88	0.20 ± .004	76.25 ± 1.76	84.30 ± 1.28
GPT-4	0.39 ± .005	**87.69** ± 1.66	**91.32** ± 1.35	**0.34** ± .004	82.36 ± 4.00	86.40 ± 3.39	0.38 ± .005	**86.26** ± 0.93	90.12 ± 0.75
Fine-tuned									
RoBERTa (Large)	0.24 ± .012	**91.97** ± 0.02	**94.85** ± 0.02	0.14 ± .005	**90.33** ± 0.25	**93.71** ± 0.21	0.22 ± .010	**91.59** ± 0.08	**94.59** ± 0.07
GPT-2 (Medium)	0.34 ± .006	90.80 ± 0.09	94.13 ± 0.07	0.12 ± .001	88.92 ± 0.38	92.82 ± 0.29	0.29 ± .005	90.36 ± 0.16	93.83 ± 0.12
T5 (Base)	0.25 ± .002	90.36 ± 0.07	93.88 ± 0.05	0.13 ± .001	88.17 ± 0.19	92.34 ± 0.15	0.23 ± .002	89.85 ± 0.08	93.52 ± 0.05
ExpInstruct	**0.37** ± .019	88.46 ± 0.83	92.05 ± 0.65	**0.30** ± .021	83.79 ± 2.74	87.58 ± 2.53	**0.35** ± .019	87.20 ± 0.69	90.96 ± 0.51

Table 5: Comparision performances between zero-shot and finetuned models with two lengthening styles. In each column, the **best** result is highlighted in bold, and the second best result is underlined.

	PP	PN	NP	NN
RLF	127	9	11	53
w/o RLF	120	16	13	51

Table 6: Confusion matrices for sample distribution. We categorized data with combinations of document and sentence labels (e.g., PP represents Positive document with Positive sentence).

Backbone Model	IAA	Reliability	S_{exp}
RoBERTa (Large)	0.41	0.57 ± .159	0.22 ± .010
GPT-2 (Medium)	0.44	0.59 ± .110	0.29 ± .005
T5 (Base)	0.42	0.48 ± .093	0.23 ± .002
GPT-4	0.43	0.79 ± .055	0.38 ± .005
ExpInstruct	0.34	0.67 ± .105	0.35 ± .019

Table 7: Detail results of Explanation Reliability for zero-shot GPT-4 and the 4 fine-tuned models (RoBERTa, GPT-2, T5, and ExpInstruct). The correlation coefficient between the reliability score and S_{exp} is 0.91.

Metric	Acc(%)		F1(%)	
Backbone Model	Zero-shot	Fine-tuned	Zero-shot	Fine-tuned
RoBERTa (Large)	93.99 ± 0.13	95.99 ± 0.51 ↑	95.82 ± 0.13	97.18 ± 0.36 ↑
GPT-2 (Medium)	89.79 ± 0.38	94.14 ± 0.16 ↑	92.72 ± 0.35	95.94 ± 0.11 ↑
T5 (Base)	92.61 ± 0.64	93.37 ± 0.06 ↑	94.84 ± 0.47	95.38 ± 0.04 ↑
LLaMA2 (13B)	71.06 ± 0.42	94.18 ± 0.63 ↑	83.04 ± 0.34	95.96 ± 0.37 ↑

Table 8: Document-level gains for LMs fine-tuned with **Lengthening**. We observe average improvement in accuracy as 7.6% and F1 score as 4.5% among the models.

	S_{exp}	Acc(%)	F1(%)
# Training Samples			
0	0.20 ± .004	76.25 ± 1.76	84.30 ± 1.28
500	0.33 ± .005	82.09 ± 0.89	86.94 ± 0.35
1,000	0.35 ± .015	85.36 ± 0.18	89.58 ± 0.26
ExpInstruct (1,600)	**0.35** ± .019	**87.20** ± 0.69	**90.96** ± 0.51
Instruction Strategy			
Instruction with SA	0.15 ± .003	86.61 ± 0.77	90.57 ± 0.70
Instruction with WIS	0.35 ± .005	72.60 ± 1.77	83.58 ± 0.84

Table 9: Results of ablation study for effects of sample size and instruction strategy.

prove both performance and explainability and present results in Table 9. Instruction with SA (ExpInstruct w/o WIS) enhances LLMs performance for SA with RLF sentences, while Instruction with WIS improves the understanding of RLF. Surprisingly, ExpInstruct gains extra benefits in both explainability and performance by combining these two strategies into one task.

Generalizability of Results We conduct two experiments to support the generalizability of our findings and show results in Table 10. 1) Fine-tune and evaluate PLMs on the 3k subset; 2) Evaluate zero-shot performance with OOD (randomly sampled 3k instances) for LLMs.

8 Discussion and Conclusion

This work sheds light on an overlooked informal style - RLF by exploring answers to two research questions. Due to the lack of existing dataset focus on RLF, we curate the **Lengthening** dataset featuring RLF grounding from 4 public datasets. We introduce **ExpInstruct** to improve the performance and explanation of LLMs for RLF. We further propose a unified approach to quantify the explainability of LMs for RLF.

Our findings uncover the expressive value of RLF from document, sentence and word levels and highlight its potential for social media content analysis, where short and informal expressions prevail. While fine-tuned PLMs achieve superior performance than zero-shot GPT-4, their understanding of RLF still needs further improvement. In addition, our results show the advantages of **ExpInstruct**, which can improve the performance and explainability of LLMs with limited samples.

Backbone Model	S_{exp}	Acc(%)	F1(%)
Fine-tuned with subset & Test on subset			
RoBERTa (Large)	0.19 ± .005	84.70 ± 0.83	89.18 ± 0.59
GPT-2 (Medium)	0.38 ± .016	79.00 ± 0.76	84.32 ± 0.59
T5 (Base)	0.12 ± .003	79.52 ± 0.76	84.97 ± 0.65
Zero-shot with OOD (randomly sample 3k)			
GPT-4	0.35 ± .008	87.55 ± 0.92	91.61 ± 0.53
LLaMA2 (13B)	0.19 ± .004	77.33 ± 0.18	87.22 ± 0.12
ExpInstruct	0.33 ± .010	90.80 ± 0.93	94.03 ± 0.60

Table 10: Comparision performances on the subset, PLMs finetuned on Lengthening and test on subset

Limitation

We acknowledge the limitations that present opportunities for future research. Firstly, human correction can improve the quality of the samples for explainable instruction. This can further improve the performance and interpretation of instructed LLaMA2. In addition, we can instruct-tune T5 and compare it with existing results. We leave this for future work.

While this study focuses on RLF in English, it is important to acknowledge that this informal style is also prevalent in other languages. For instance, in Spanish, words like "graciaaas" or "holaaa!!!!" are used to convey friendliness or emphasis. In Romanian, words like "daaaa" (yes) are repeated to show strong agreement, and "minunat!!!" (wonderful!!!) to express amazement or excitement. RLFs are also commonly used in daily communications in other languages such as Chinese and Arabic. Although this paper uses datasets in English, our methodologies, including dataset generation, the ExpInstruct framework, and the unified approach for explainability evaluation can be easily transferred to other languages.

Acknowledgements

This work was supported by the National Science Foundation awards IIS-1546480 and ITE-2137846. We also thank Andrew Schneider for his valuable contributions to our project.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Farman Ali, Daehan Kwak, Pervez Khan, Shaker El-Sappagh, Amjad Ali, Sana Ullah, Kye Hyun Kim, and Kyung-Sup Kwak. 2019. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, 174:27–42.

Abdullah Aljebreen, Weiyi Meng, and Eduard Dragut. 2021. Segmentation of tweets with urls and its applications to sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12480–12488.

Rami Aly, Xingjian Shi, Kaixiang Lin, Aston Zhang, and Andrew Wilson. 2023. Automated few-shot classification with instruction-finetuned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2414–2432, Singapore. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Francesco Barbieri, Miguel Ballesteros, Francesco Ronzano, and Horacio Saggion. 2018. Multimodal emoji prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 679–686, New Orleans, Louisiana. Association for Computational Linguistics.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2022. What do llms know about financial markets? a case study on reddit market sentiment analysis.

Eduard Dragut and Christiane Fellbaum. 2014. The role of adverbs in sentiment analysis. In *Proceedings of*

Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014), pages 38–41.

Eduard Dragut, Hong Wang, Clement Yu, Prasad Sistla, and Weiyi Meng. 2012. Polarity consistency checking for sentiment dictionaries. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 997–1005, Jeju Island, Korea. Association for Computational Linguistics.

Eduard C. Dragut, Hong Wang, Prasad Sistla, Clement Yu, and Weiyi Meng. 2015. Polarity consistency checking for domain independent sentiment dictionaries. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):838–851.

Eduard C. Dragut, Clement Yu, Prasad Sistla, and Weiyi Meng. 2010. Construction of a sentimental word dictionary. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, page 1761–1764, New York, NY, USA. Association for Computing Machinery.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Kyle Gorman, Christo Kirov, Brian Roark, and Richard Sproat. 2021. Structured abbreviation expansion in context. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 995–1005, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tyler J. Gray, Christopher M. Danforth, and Peter Sheridan Dodds. 2020. Hahahahaha, duuuuude, yeeessss!: A two-parameter characterization of stretchable words and the dynamics of mistypings and misspellings. *PLOS ONE*, 15(5):e0232938.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Lihong He, Chao Han, Arjun Mukherjee, Zoran Obradovic, and Eduard Dragut. 2019. On the dynamics of user engagement in news comment media. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10.

Lihong He, Chen Shen, Arjun Mukherjee, Slobodan Vucetic, and Eduard Dragut. 2021. Cannot predict comment volume of a news article before (a few) users read it. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):173–184.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Marjan Hosseinia, Eduard Dragut, Dainis Boumber, and Arjun Mukherjee. 2021. On the usefulness of personality traits in opinion-oriented tasks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 547–556, Held Online. INCOMA Ltd.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Kango Iwama and Yoshinobu Kano. 2018. Japanese advertising slogan generator using case frame and word vector. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 197–198, Tilburg University, The Netherlands. Association for Computational Linguistics.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the tweebank corpus on named entity recognition and building nlp models for social media analysis. *In Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*.

Yoram M. Kalman and Darren Gergle. 2014. Letter repetitions in computer-mediated communication: A unique link between spoken and online language. *Computers in Human Behavior*, 34:187–193.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

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.

Jiwei Li. 2020. Hotel review dataset. Access date: 2023-10-11.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Understanding neural networks through representation erasure.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *arXiv preprint arXiv:2401.01523*.

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

Keming Lu, Xiaoman Pan, Kaiqiang Song, Hongming Zhang, Dong Yu, and Jianshu Chen. 2023. PIVOINE: Instruction tuning for open-world entity profiling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15108–15127, Singapore. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Shotaro Misawa, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. 2020. Distinctive slogan generation with reconstruction. In *Proceedings of Workshop on Natural Language Processing in E-Commerce*, pages 87–97, Barcelona, Spain. Association for Computational Linguistics.

Huy Nguyen and Minh-Le Nguyen. 2017. A deep neural architecture for sentence-level sentiment classification in twitter social networking. *CoRR*, abs/1706.08032.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report.

Letian Peng, Zilong Wang, Hang Liu, Zihan Wang, and Jingbo Shang. 2023. Emojilm: Modeling the new emoji language. *arXiv preprint arXiv:2311.01751*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. Association for Computational Linguistics.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models “understand” language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jens Reelfs, Timon Mohaupt, Sandipan Sikdar, Markus Strohmaier, and Oliver Hohlfeld. 2022. Interpreting emoji with emoji. In *Proceedings of the Fifth International Workshop on Emoji Understanding and Applications in Social Media*, pages 1–10, Seattle, Washington, USA. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ""why should I trust you?"": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

Tyler Joseph Schnoebelen, P. Eckert, D. Jurafsky, C. Potts, and J. R. Rickford. 2012. *Emotions are relational: Positioning and the use of affective linguistic resources*. Ph.d. thesis, Stanford University. Submitted to the Department of Linguistics.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? generating explanations for visual semantic role labelling in memes. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

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

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Fan Yang, Eduard Dragut, and Arjun Mukherjee. 2020. Predicting personal opinion on future events with fingerprints. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1802–1807, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yelp. 2021. Yelp dataset. Yelp. Retrieved Feb 16, 2021.

Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4031–4047, Singapore. Association for Computational Linguistics.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023a. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023b. Sentiment analysis in the era of large language models: A reality check.

Zhexin Zhang, Jiale Cheng, Hao Sun, Jiawen Deng, and Minlie Huang. 2023c. InstructSafety: A unified framework for building multidimensional and explainable safety detector through instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10421–10436, Singapore. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert.

Yangqiaoyu Zhou, Yiming Zhang, and Chenhao Tan. 2023. FLamE: Few-shot learning from natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6743–6763, Toronto, Canada. Association for Computational Linguistics.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. 2024. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts.

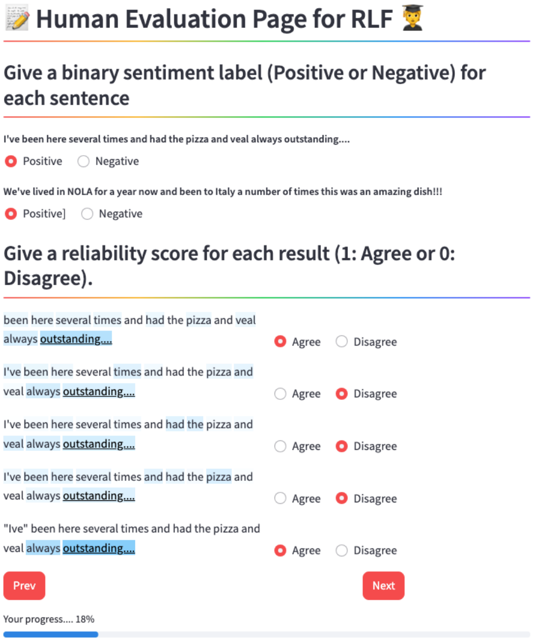
A Dataset Details

A.1 Algorithm for Identifying RLF

Sentence Segmentation and Merging. Utilizing open-source regular expressions (regex)[1], we segment documents into individual sentences. Subsequently, to prevent the formation of trivial fragments, we merge sentences comprising fewer than three words. This approach ensures we retain meaningful linguistic structures.

Model	#Dataset (k)	#train (k)	#val (k)	#test (k)	data split	experiment
PLMs	850	595.2	84.8	170	7 : 1 : 2	3 runs
ExpInstruct	3	1.6	0.4	1.0	custom	3-fold

Table 11: Details of dataset split.



📝 Human Evaluation Page for RLF 👨‍🎓

Give a binary sentiment label (Positive or Negative) for each sentence

I've been here several times and had the pizza and veal always outstanding....

● Positive ○ Negative

We've lived in NOLA for a year now and been to Italy a number of times this was an amazing dish!!!

● Positive] ○ Negative

Give a reliability score for each result (1: Agree or 0: Disagree).

been here several times and had the pizza and veal always outstanding.... ● Agree ○ Disagree

I've been here several times and had the pizza and veal always outstanding.... ○ Agree ● Disagree

I've been here several times and had the pizza and veal always outstanding.... ○ Agree ● Disagree

I've been here several times and had the pizza and veal always outstanding.... ○ Agree ● Disagree

"Ive" been here several times and had the pizza and veal always outstanding.... ● Agree ○ Disagree

Your progress.... 18%

Figure 5: Our customized user interface for human evaluation. Annotators are asked to do two tasks: annotation for sentiment labels and explanation reliability.

Find Root Word for RLF. The process involves determining the root words of these lengthened forms using a reduced method in (Kalman and Gergle, 2014) grounded in American English[2]. For instance, variations like 'loooove' and 'loooovvve' have the generalized forms of 'lo+ve' and 'lo+v+e', and both come from the root word 'love'. We further refine our dataset by retaining only those instances where the frequency of the generalized form exceeds 100 occurrences, thereby filtering out less common variations.

POS Tagging for RLF Words. Initially, we substitute each lengthened word in the sentence with its corresponding root word. Following this, we employ TweebankNLP (Jiang et al., 2022) to ascertain the part-of-speech (POS) tag of the root word. This POS tag(Dragut and Fellbaum, 2014) is then attributed to the respective lengthened word. We present examples of normalized forms by lengthening style and POS tag in Table 12.

Pairing RLF Sentence with w/o RLF Sentence For every sentence identified with RLF words, we select a corresponding w/o RLF sentence from the same document (provided it contains two or more sentences) to serve as a control sample. This allows for a comparative analysis of them. We attach the document's overall sentiment label to individual sentences extracted from it.

Dataset Balancing We balance the data distribution by applying downsampling to dominant domains and lengthening styles. Specifically, we strategically sample 20% of sentences with repetitive letters, 8% of those with ellipses, and all other repetitive punctuation. Additionally, to avoid data imbalance due to specific domains or generalized forms, we downsample the most prevalent ones, ensuring a more uniform distribution across the dataset.

A.2 Dataset Split

The Lengthening dataset consists of 850k samples as shown in Table 1. For our experiments with LLMs, we randomly sampled a subset of 3000 instances and ran experiments with 3-fold cross-validation strategy. In each fold, 1.6k instances were used for training, with the remaining data for validation and testing. The details of the data split are as shown in Table 11.

B Implementation Details

B.1 Details of Model Parameters

All models are trained with a learning rate of 2e-5, a weight decay of 0.01, a maximum gradient norm of 1.0, and utilized gradient accumulation over 4 steps. We instruct-tune LLaMA2-13B-chat-hf[3] with the subset samples for 5 epochs using a batch size of 1. It was set to a learning rate of 2e-4, with a reduced weight decay of 0.001 and a maximum gradient norm of 0.3, implementing a single

[1] https://stackoverflow.com/questions/4576077/how-can-i-split-a-text-into-sentences

[2] https://pyenchant.github.io/pyenchant/tutorial.html

[3] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

Lengthening Style	POS Tag	Ratio (%)	Examples of normalized forms
Letter Repetitive	ADV	14.41	’so+’, ’wa+y’, ’way+’, ’reall+y’, ’wa+y+’
Letter Repetitive	INTJ	6.67	’hmm+’, ’ah+’, ’oh+’, ’um+’, ’aw+’
Letter Repetitive	others	3.20	’all+’, ’to+’, ’go+d’, ’bu+t’, ’al+’
Letter Repetitive	VERB	3.19	’lo+ve’, ’love+’, ’lo+ved’, ’recomm+end’, ’lo+ves’
Letter Repetitive	ADJ	3.06	’lo+ng’, ’hu+ge’, ’slo+w’, ’long+’, ’litt+le’
Letter Repetitive	NOUN	1.71	’boo+k’, ’go+d’, ’a+s’, ’wa+y’, ’boo+’
Letter Repetitive	PRON	0.26	’you+’, ’me+’, ’who+’, ’it+’, ’sh+’
Punctuation Repetitive	NOUN	29.24	’book!+’, ’read!+’, ’series!+’, ’place!+’, ’product!+’
Punctuation Repetitive	ADJ	10.38	’amazing!+’, ’awesome!+’, ’great!+’, ’good!+’
Punctuation Repetitive	ADV	9.07	’again!+’, ’ever!+’, ’back!+’, ’here!+’, ’down!+’
Punctuation Repetitive	PRON	6.87	’it!+’, ’you!+’, ’it...+’, ’this!+’, ’them!+’
Punctuation Repetitive	VERB	6.59	’read!+’, ’work!+’, ’sucks!+’, ’had!+’, ’go!+’
Punctuation Repetitive	others	4.03	’for!+’, ’amazon!+’, ’etc...+’, ’it!+’
Punctuation Repetitive	INTJ	1.31	’ah+’, ’please!+’, ’um+’, ’yes!+’, ’aw+’

Table 12: Examples of normalized forms by lengthening style and POS tag.

step for gradient accumulation. Distinctively, 4-bit quantization was enabled for LLaMA2, and the compute data type was set to Float16. Moreover, the LoRA framework was integrated, set with a rank size of 64, an alpha value of 16, and a dropout probability of 0.1. For LLaMA2, we set the temperature at 0.2, the repetition penalty at 1.4, and the maximum length to the length of the prompt plus 10 tokens. This configuration was designed to limit the size of the output token and prevent excessively long responses, thereby conserving response time and enhancing computational efficiency.

For the GPT-4[4] API, we configured the temperature setting at 0.2 and established a maximum token size limit of 5,000, with other parameters remaining at default values.

We fine-tune 3 PLMs with backbones as RoBERTa (Large) [5], also referred to as SIEBERT (Hartmann et al., 2023), GPT-2 (Medium)[6] and T5 (Base) [7].

B.2 Analysis of RLF Styles

In this section, we evaluate the performance of various models across two distinct styles of RLF, with results detailed in Table 5. We observe the Punctuation Repetitive style consistently achieve higher scores across three evaluation metrics and with both zero-shot and fine-tuned models, suggesting that this style is stronger for sentiment expression.

C Details of Human Evaluation

We hire 3 graduate students as annotators for human evaluation. All annotators have fluent English levels. Specifically, we sample 200 instances from the subset dataset and ask annotators to conduct two annotation tasks. We guarantee annotators receive fair wages of 20$ per hour.

Annotation for Sentiment Label Give a sentence (RLF or w/o RLF). Annotators need to give a binary sentiment label (1: Positive, 0: Negative).

Annotation for Explanation Reliability We disorder and list WIS results for an RLF sentence from the 5 LMs (3 fine-tuned PLMs, ExpInstruct, and zero-shot GPT-4). Annotators need to give the reliability score for each result (1: Agree, 0: Disagree).

We customized our annotation page with streamlit[8] and present a case in Figure 5. Source code and sample data for this page can be found with our project link.

[4] https://platform.openai.com/playground?mode=chat&model=gpt-4

[5] https://huggingface.co/siebert/sentiment-roberta-large-english

[6] https://huggingface.co/michelecafagna26/gpt2-medium-finetuned-sst2-sentiment

[7] https://huggingface.co/mrm8488/t5-base-finetuned-imdb-sentiment

[8] https://streamlit.io/install