Popular Quoting Tweet Generation via Auto-Response Augmentation

Anonymous ACL submission

Abstract

A quoting tweet allows users to share others' content while adding their comments. To help users write a quoting tweet with better public engagement, we study the task of popular quoting tweet generation. The focus is to generate quoting tweets with higher popularity reflected by more likes, replies, and retweets. While large language models (LLMs) showed exceptional language generation capabilities, limited work has examined how LLMs can learn the popularity of text to engage the public better. Consequently, we propose a novel Response-augmented Popularity-Aligned Language Model (RaPALM) to align language generation to popularity by incorporating insights from augmented automatic responses. Here, we employ the Proximal Policy Optimization (PPO) framework with a dual-reward mechanism to jointly explore popularity in quoting tweet generation. The experiments on two newly gathered datasets of quoting tweets for external links or others' tweets show that RaPALM exhibits state-of-the-art results.[1]

1 Introduction

A *quoting tweet* allows users to share external links or other users' tweets while adding their comments. Its purpose is to enhance the visibility of the source message, beneficial to various applications, e.g., media broadcasts, advertisements, and so forth (Lin et al., 2023). A popular quoting tweet can prompt public readers to actively engage in the discussions. It can broaden the dissemination of the source message and incite a more dynamic discourse and exchange of viewpoints among users. Previous work showed that the wording of tweets could substantially impact *popularity*, reflected by user replies, retweets, and likes (Tan et al., 2014).

Nevertheless, many users are not good at writing popular quoting tweets. To help them better engage the public for meaningful interactions, we study a novel task of how NLP models can learn to generate a popular quoting tweet given a source message of an external link or other users' tweets.

Source Message: ChatGPT-A Silver Bullet for Your Customer Support Org? Language models like ChatGPT can write blog posts, hold conversations, and even pass the bar.
A Popular Quoting Tweet (manually written): Will ChatGPT replace customer support teams? At @users, ***we've already deployed language models like ChatGPT*** to help support orgs like ... at scale. ***Learn what this means for you*** and how your company can stay ahead.
LLaMA2-Chat: Pondering the future of #customersupport: Will #ChatGPT be the silver bullet for orgs? #AI #languagemodel
ChatGPT: Revolutionizing Customer Support with ChatGPT! Discover how language models like ChatGPT are not just conversing and blogging, but also acing legal tests. Is this the future of customer service? #ChatGPT #CustomerServiceInnovation #AIRevolution
RaPALM: ***Just set up my ChatGPT*** and I'm blown away by its capabilities! ***Just learn and try it on your customer support team***. Will it replace human agents? Maybe not, but it's definitely a game-changer for customer service. #ChatGPT

Table 1: A sample source message about "*Capabilities of ChatGPT on Customer Service*" and a manually-written popular quoting tweet on the top. Below are three quoting tweets generated by different LLMs. The same colors in purple and red indicate similar meanings.

Despite the recent advances of LLMs in language generation (Wei et al., 2021; Ouyang et al., 2022b), the mainstream research focuses on the writing itself, yet limited work concerns the public readers' reactions to the text. For this reason, existing models cannot effectively model the text's popularity, which reflects the potential to draw public engagement. To illustrate this point, Table 1 shows a sample source message of news followed by the manually written and automatically quoting tweets. We observe that LLaMA2-chat (Touvron et al., 2023) and ChatGPT (Ouyang et al., 2022c) simply summarize the news without incorporating any additional insights, thus unlikely to draw engagement. On the contrary, the manually written reference is rich in original thoughts and opinions.

[1]Our code and dataset are available at https://anonymous.4open.science/r/RaPALM-14AA/.

Viewing the limitation of LLMs in popularity learning, we propose a novel response-augmented popularity-aligned language model (RaPALM). RaPALM relates quoting tweet's language to popularity by employing LLMs to predict possible reader responses, which serve as a mirror to reflect public reactions for potential engagement measurements. Augmented by these (auto-)responses, RaPALM is trained to align the quoting tweet writing to popularity measure via reinforcement learning (RL).

Concretely, we first gather multiple LLM-generated auto-responses and select those that best match the source message with a consistency matching method. Then, we feed a source message with its selected responses into RaPALM to generate multiple quoting tweets. Next, we optimize RaPALM's training process with the PPO framework (Schulman et al., 2017) with a novel dual-reward design. Here, one reward is to predict popularity trained with popular quoting tweets in positive-negative sample pairs. The other measures consistency to auto-responses to align with public reactions. Finally, we develop a reward ranking and sampling method to select high-reward training examples to improve training effectiveness.

To the best of our knowledge, *RaPALM is the first model to utilize LLM-predicted auto-responses for popularity-aligned language generation*. By learning from these potential responses, RaPALM can effectively generate popular quotable tweets that helpfully draw engagement. For example, as illustrated in Table 1, the output of RaPALM is rich in eye-catching viewpoints, such as "*blown away by its capabilities*" and "*just learn and try it*."

As a pilot study on popular quoting tweet generation, we benchmark the task with two datasets. One is named QuoteLink containing tweets quoting external links and the other QuoteTweet quoting other users' tweets. The two datasets contain 70K pairs of positive-negative samples; each pair of tweets quotes the same source and is from the same author, yet one (the positive sample) is more popular.

We further experiment with the two datasets. The main results first show that RaPALM outperforms all comparison models in both automatic measure and human evaluation. For example, RaPALM achieves 23.26 Rouge-1, compared to 20.94 from ChatGLM3. Then, the ablation study implies the positive contributions of varying RaPALM modules. Next, quantitative analyses show the effectiveness of RaPALM in varying scenarios. After that, we conduct a case study to interpret why RaPALM can perform better. Lastly, we analyze the quoting tweet wording from four aspects to examine what affects popularity to inspire future work.

In summary, our contributions are threefold:

• We present the first study on popular quoting tweet generation with two large-scale datasets.

• We propose RaPALM with dual-reward RL to exploit auto-responses to reflect public reactions for aligning language generation to popularity.

• We extensively experiment with RaPALM and demonstrate its state-of-the-art (SOTA) performance in generating popular quoting tweets.

2 Related Work

Quoting Tweet Generation. As this is a newly proposed task, here we discuss two potential lines of methods that can apply to our task: summarization and headline generation. The summarization methods (Phang et al., 2022; Lewis et al., 2020) aim to extract the salient information from the source text. The headline generation (Kanungo et al., 2021; Zhang et al., 2020) task aims to create a headline to summarize or quote the source's content. However, most methods focused on the writing without considering the popularity factors for further public engagements on social media.

Our work is related to language generation in a broader scope. The emergence of LLMs has substantially advanced this field, especially in the zero-shot domain. Taking recent advances in LLMs, many studies have examined how to align language models with human feedback. For example, ChatGPT, a closely related model to InstructGPT (Ouyang et al., 2022b), is specifically trained to follow human instructions. LLaMA2-chat (Touvron et al., 2023) is an open-source language model that demonstrates SOTA performance in conversational abilities. Our RaPALM explores aligning the language model with popularity for quoting tweet generation, which has not been explored previously.

Response Augmentation. Our work is also inspired by previous work enriching context with augmented responses to provide readers' views and help NLP models use languages. Xu and Li (2022) borrowed human senses by retrieving responses for social media multimodal classification. Niu et al. (2023) incorporated responses to supplement image features for image aesthetics assessment. Liu et al. (2023) employed human responses for humor detection in short-form videos. However, previous related work mainly relies on existing responses,

Datasets	Pair Number			Avg. Token Number			Avg. Popularity Gap		
	Train	Valid	Test	Src	Pop	UnPop	Like	Reply	Retweet
QuoteLink	18,969	6,323	6,323	186.7	135.1	158.6	299.4	14.1	53.7
QuoteTweet	21,892	7,298	7,298	156.1	92.9	118.9	158.1	15.5	57.3

Table 2: Statistical of popular quoting tweets datasets. We report the two datasets' pair number, average token number, and popularity gap. "Avg. Popularity Gap" refers to the average difference in "like", "retweet" or "reply". For instance, a "Like" value of 299.4 indicates that, on average, Tweet A receives 299.4 more likes than Tweet B.

which cannot be applied in scenarios without human responses. On the contrary, we make the first efforts to utilize LLMs to simulate potential user responses automatically and enable language generation models to gain a better sense of popularity.

Popularity Analysis. Our work is further related to popularity prediction on social media, where users express their preferences by replying, liking, or retweeting behavior. The count of such behavior is usually adopted as the popularity indicator. Tan et al. (2014) analyzed the effect of wording on tweet propagation. Lamprinidis et al. (2018) used a multi-task GRU network to predict headline popularity. Kano et al. (2018) employed such popularity measure to supervise extractive summarization distantly. Gao et al. (2020) leveraged social media feedback data to build a large-scale dataset to predict popularity. However, none of them explores how to engage the popularity factors in language generation, which we will extensively explore.

3 Quoting Tweet Datasets

We collected large-scale data from Twitter for our popular quoting tweet generation task. Based on the source of the quotes, we separated the data into two distinct datasets: ***QuoteLink*** and ***QuoteTweet***, specifically for writing quoting tweets for external links and internal tweets, respectively.

Data Collection. We first downloaded the general Twitter streams from September 2018 to September 2019 from Nguyen et al. (2020). Then, we removed duplicate users and shortlisted the tweets from users with over 10,000 followers. The reason for that is to choose tweets with a specific degree of visibility to impartially measure popularity. Subsequently, we separate selected tweets by the types of source messages in two datasets. One is to quote an external link attached at the end of the text, which we used for the QuoteLink dataset. The other contains tweets quoting other users' tweets corresponding to the QuoteTweet dataset. After that, we gathered the content of these tweets with source messages and measured the number of likes, replies, and retweets to reflect popularity. Finally, we retained the tweet text in English and removed irrelevant fields, such as images and videos.

Tweet Pair Construction. Given that popularity is a subjective concept, we follow Tan et al. (2014) to train models using positive-negative quoting tweet pairs. A tweet pair is from the same user quoting the same source while one (positive sample) is more popular than the other (negative sample). A positive-negative pair is labeled as (Tweet A, Tweet B). To construct such pairs, we implemented four steps following Tan et al. (2014): 1) Tweets A and B must be from the same author and contain the same source message. 2) Tweet A must have at least 10 more likes, replies, or retweets than Tweet B. 3) The posting time interval between Tweet A and Tweet B must be less than 12 hours. 4) As suggested by (Tan et al., 2014), we used SimCSE (Gao et al., 2021) to measure the semantic similarity of the tweet pair and removed too-similar pairs (with over 0.53 similarity) for easier comparison. For model training and testing, we randomly split each of the datasets into training (70%), validation (15%), and test (15%) sets.

Data Analysis. Table 2 shows our two datasets' pair numbers, average token number, and popularity gap. We observe that in the QuoteLink dataset, the average length of tweets is generally longer than in the QuoteTweet dataset. It indicates that users possibly tend to add more words and detailed information when quoting external links. For the popularity gap, popular quoting tweets (positive samples) in both datasets have significantly higher likes, replies, and retweets than unpopular ones (negative). It demonstrates the datasets will allow a meaningful comparison for popularity learning.

4 RaPALM Framework

RaPALM overview. To begin with, we describe our datasets as $D = \{s^i, t_u^i, t_p^i\}_{i=1}^N$, where s^i

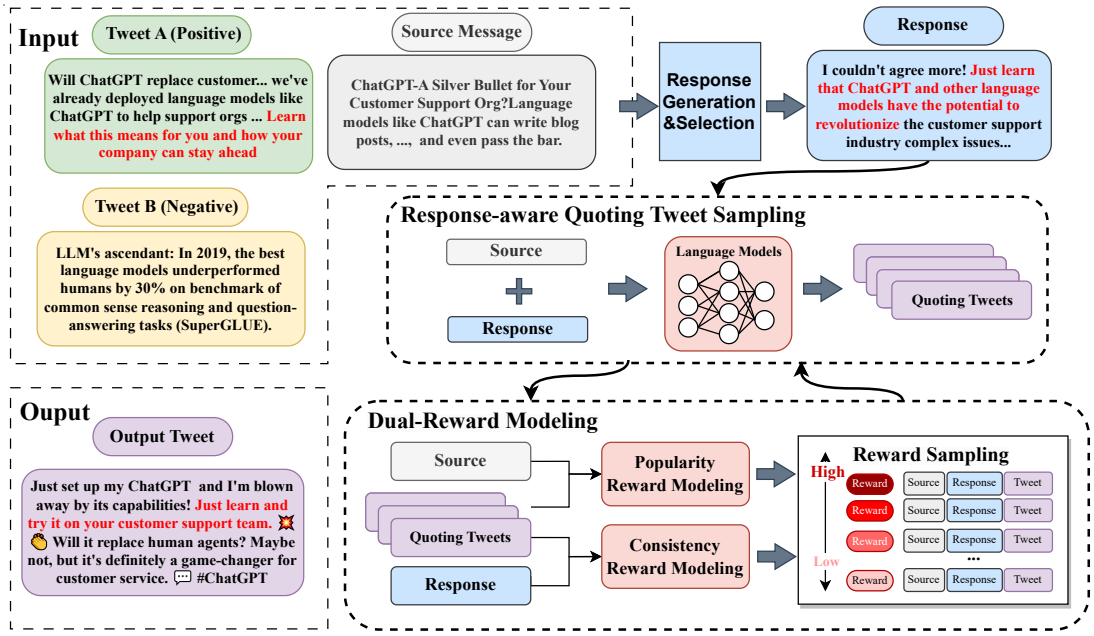


Figure 1: The workflow of RaPALM is outlined as follows: the first step involves **generating potential public responses (4.1)** based on source massages and selecting them based on semantic consistency to the source. In the second step, the selected response is leveraged to help **generate possible quoting tweets (4.2)**. Then, the designed **dual-reward modeling (4.3)** method calculates the rewards for these generated quoting tweets. Finally, the data is chosen for optimization through the **data sampling method (4.4)**.

stands for the source message, which could be either an external link or a general tweet. t_u^i and t_p^i represent the unpopular and popular tweets of the same user quoting s^i, and N is the pair number. The goal of RaPALM is to generate a popular quoting tweet t_p based on the source s (we omit the index i for better illustration). Its workflow is depicted in Figure 1, which includes four components: auto-response generation and selection, response-aware quoting tweet sampling, dual-reward modeling, data sampling and learning.

4.1 Auto-Response Generation and Selection

Previous work has incorporated human response into language models, allowing them to possess a human-like sense. However, these methods rely on the existing responses. When we generate quoting tweets, we face the challenge that public reactions have not yet formed, leaving us without existing responses to refer to. To address this issue, we simulate potential public reactions, enabling our language model to effectively generate popular quoting tweets, even without actual responses. We first prompt the LLM to sample different responses. Then, to ensure that the generated response is consistent with the source message, we calculate the semantic similarity between them. After that, we rank the responses based on their similarity, and the top-ranking response is selected to help generate quoting tweets. The process can be formulated as follows:

$$\begin{aligned} R_{sampled} &= LLM(s) \\ resp &= MaxSim(R_{sampled}, s) \end{aligned} \tag{1}$$

where the SimCSE-measured cosine similarity is used to calculate the semantic similarity. $MaxSim$ function finds the response in $R_{sampled}$ that is most similar to s.

4.2 Response-aware Quoting Tweet Sampling

After obtaining the human response, we incorporated it into the process of quoting tweet generation. Firstly, we experimented with various prompt templates to merge the source message and the generated response. Ultimately, we adopted the most effective prompt template: "Given the news [source] and potential public reaction [human response], create a quoting tweet that highlights the main point of the news while capturing the public's response." Then, we controlled the temperature parameter α to sample multiple quoting tweets. The process can be described as:

$$T_{sampled} = \pi^{\alpha}(\phi)(prompt[s, resp]) \tag{2}$$

where $T_{sampled} = \{t_1, t_2, ...t_k\}$ is the sampled quoting tweets and k denotes the number of samples. The function $prompt[.]$ concatenates the

source message s and response r according to the template. π is an LLM the PPO aims to optimize.

4.3 Dual-Reward Modelling

Although LLMs can generate quoting tweets, they have not considered the popularity factor. Inspired by RLHF (Ouyang et al., 2022a), we utilize the PPO framework and design dual-reward modeling to align LLM with popularity. The dual-reward model consists of popularity reward modeling and consistency reward modeling.

Popularity reward modeling primarily assesses whether social media users will engage with the generated tweet. Specifically, it outputs a scalar reward by taking the generated quoting tweet and its corresponding source message as input. The loss function for the popular reward model is:

$$\mathcal{L}_{RM}^{pop}(\theta) = -E_{(s,t_u,t_p)\sim D} [\log(\sigma(r_\theta^{pop}(s,t_u) - r_\theta^{pop}(s,t_p)))] \quad (3)$$

where θ is the training parameters of the popular reward model. $r_\theta^{pop}(s,t)$ is the scalar output of the reward model for source s and tweet t.

Consistency reward modeling evaluates if the generated quoting tweet aligns with the generated human response. Our objective is that these generated tweets could capture the essence of the human response. To achieve this, we measure the similarity between the response and the quoting tweet using unsupervised SimCSE, denoted as $r^{cons}(s,t)$. The overall reward $r(s,t)$ is the sum of the two rewards:

$$r(s,t) = r_\theta^{pop}(s,t) + r^{cons}(s,t) \quad (4)$$

4.4 Data Sampling and Learning

Training PPO typically requires high-quality data, such as human feedback provided by experts. However, our dataset, being automatically collected from social media, cannot guarantee that each training data is high quality.

Inspired by Dong et al. (2023), who chose to fine-tune their model using examples with high rewards. We collect multiple pairs of reward-source-tweet (r,s,t) via the above methods. This provides us with a selective approach to extract only high-reward samples for subsequent PPO training. Specifically, we rank the collected pairs and select the top k percent of samples with the highest rewards as our sampled training datasets D_{RL}. Our PPO training function can be defined as:

$$\mathcal{L}_{RL} = -E_{(r,s,t)\sim D_{RL}} r(s,t) \quad (5)$$

5 Experiment Setup

5.1 Model Settings

We use LLaMA2 (Touvron et al., 2023) (LLaMA2-chat-7b version is adopted in all experiments) as our auto-response generation model, which is not involved in training and is only used for sampling responses. For each source message, we sample 5 responses and further perform semantic matching.

During the model training phase, we also employ LLaMA2 as our quoting tweets generation policy. We did not fine-tune this model, as doing so could potentially decrease its performance. This might be because the model has already been adjusted through instruction tuning, making such word-level adjustments unnecessary. The max length of the generated tweet is set to 150, sampling number k and temperature α is set to 5 and 0.6. We use GPT-2 as our popularity reward model and train it on our quoting tweet pairs. Additionally, we directly utilize unsupervised SimCSE as our consistency reward model. For the PPO training process, we set the learning rate to 2e-5, the batch size to 4, and the training epochs to 3. For more training details, please refer to our code.

5.2 Evaluation Metrics

For *Automatic Evaluation*, we compare generated quoting tweets with popular tweets and evaluate the output quality with metrics of ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), NIST (Lin and Hovy, 2003) and BertScore (Zhang* et al., 2020).

For *Human Evaluations*, we invited human raters with NLP backgrounds to select preference between the generated tweet of different models considering two dimensions: *consistency* of a generated tweet to the source message, and *popularity* of the tweet that its the potential to engage the public.

5.3 Baselines and Comparison

We adopt summarization and headline generation models for comparison. For summarization models, we utilized SOTA summarizers, 1) PEGASUS-X (Phang et al., 2022) and 2) BART-Summary (Lewis et al., 2020). Additionally, we used T5 (Chung et al., 2022) to generate headlines 3) T5-Headline and quoting tweet 4) T5-Tweet. For aligned large language models, our comparisons included 5) ChatGLM3-6B (Du et al., 2022) and 6) LLaMA2 (Touvron et al., 2023). We also employ the selected responses in our method 7) LLaMA2-Response as a baseline.

Models	QuoteLink					QuoteTweet				
	R-1	R-L	BLEU	NIST	BertS	R-1	R-L	BLEU	NIST	BertS
PEGASUS-X	16.90	13.37	10.87	0.37	84.37	9.25	7.26	5.92	0.19	81.61
Bart-Summary	17.45	12.84	12.08	0.38	81.21	10.53	7.95	5.88	0.21	80.23
T5-Headline	16.74	13.36	12.50	0.43	82.94	9.49	7.75	5.63	0.19	80.64
T5-Tweet	12.35	10.57	8.85	0.34	82.17	6.02	5.23	4.03	0.14	80.33
LLaMA2-Response	17.21	11.81	12.30	0.56	83.12	11.37	8.03	8.46	0.37	80.43
LLaMA2	19.61	14.18	14.57	0.66	83.55	11.59	8.52	8.66	0.37	81.27
ChatGLM3	20.94	15.49	15.46	0.69	84.11	11.91	8.84	9.21	0.39	82.32
RaPALM	**23.26**	**15.98**	16.33	**0.74**	**84.71**	**14.18**	**10.69**	**11.98**	0.51	**83.32**
-w/o Response Augmentation	20.79	14.78	15.03	0.63	83.12	12.01	9.11	9.34	0.33	82.07
-w/o Dual-Reward Modeling	21.37	14.34	16.21	0.72	83.78	14.01	10.12	11.67	**0.53**	81.79
-w/o Reward Sampling	22.65	15.67	**16.51**	0.72	84.59	13.93	10.61	11.77	0.43	81.84

Table 3: Main comparison results and ablation result on QuoteLink and QuoteTweet. We report the evaluation metrics R-1(Rouge-1), R-L(Rouge-L), BLEU, NIST, and BertScore (BertS). Our model achieves the best results in all evaluation methods (bold and underlined), and the performance gain is significant for all comparison models (measured by paired t-test with p-value<0.05).

6 Experimental Results

We present the results of our automatic evaluation in Section 6.1 and those of the human evaluation in Section 6.2. The ablation study examining the impact of various components is detailed in Section 6.3. Quantitative analysis of how different parameters affect outcomes is referred to in Section 6.4. Additionally, a case study is discussed in Section 6.5, and an analysis of the wording in the generated tweets can be found in Section 6.6.

6.1 Main Comparison Results

Table 3 (top) shows the result. We draw the following observations: 1) Generating tweets to quote a user's tweet is more challenging than quoting an external link, possibly because user tweets are shorter and lack sufficient information. Our RaPALM can enhance the effectiveness of quoting tweets through the method of response augmentation. 2) Using summary models or headline generation models for generating quoting tweets results in poor performance. The reason could be these models typically focus on summarizing information rather than expressing their own viewpoints. Additionally, the responses generated by LLaMA2 did not perform well, as they often included content unrelated to the source message. 3) ChatGLM3 and LLaMA2 have shown promising results in generating tweets. These LLMs leverage extensive training data and contextual understanding to produce coherent, contextually relevant, and engaging tweets. 4) Built on these LLMs, our model has achieved better results. For example, RaPALM achieves 23.26 and 14.18 Rouge-1 in the two datasets, compared to 19.61 and 11.59 from LLaMA2. The promising result indicates the effectiveness of our response-augmented and popularity-aligned mechanism.

6.2 Human Evaluation

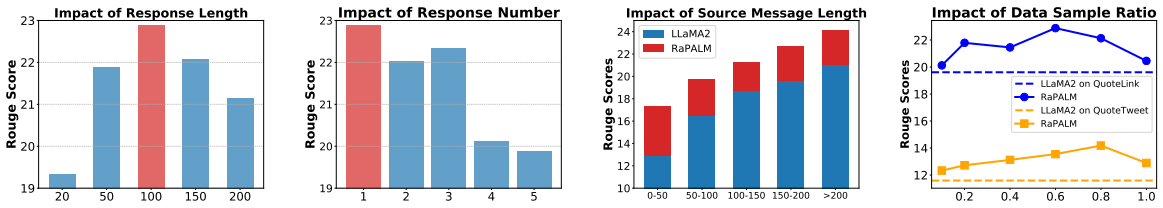
We conduct manual pair-wise evaluations to assess the consistency and popularity of the top-performing model (LLaMA2), our proposed model, and its ablation without auto-response augmentation (-w/o response). The results are shown in Table 4. We observed that incorporating responses into the model helps improve the consistency and popularity of the tweets generated, indicating that providing the model with public reactions is effective. After training with our framework, our generated tweets received more popularity than LLaMA2 and maintained consistency with the original context.

6.3 Ablation Study

The above results show the overall superiority of our model. To further investigate the effects of its components, we conduct an ablation study with response augmentation, dual-reward modeling and reward sampling.As can be seen in Table 3 (bottom), all components contribute positively to the model's performance. Notably, the model's performance declines the most when responses are reduced, indicating that the public reactions enhance the model's effectiveness.

6.4 Quantitative Analysis

We conduct quantitative analysis to better study our model. We quantify the response length, response number, and source length, and sample ratio k to examine how they affect performance.



(a) (b) (c) (d)

Figure 2: Quantitative analysis results on a) response lengths, b) the number of responses, c) source messages length, and d) data sample ratio, where rouge-1 score is adopted as the evaluation metric.

Choice %	RaPALM *vs* RaPALM$_{\text{w/o response}}$		
	RaPALM	**w/o response**	*Kappa*
Cons.	**62.3**	37.7	0.382
Pop.	**66.0**	34.0	0.434
Choice %	**RaPALM *vs* LLaMA2**		
	RaPALM	**LLaMA2**	*Kappa*
Cons.	**65.3**	34.7	0.388
Pop.	**68.3**	31.7	0.379

Table 4: Human Evaluation w.r.t. consistency and popularity. The score is the percentage that the proposed model wins against its competitor. Kappa denotes Fleiss' Kappa (Fleiss, 1971), which indicates all of our evaluation annotations reach a fair or moderate agreement.

Varying Response Length and Number. The first parameter analysis concerns the response length. As shown in Figure 2(a), the score first increases, and peaks at length 100, then decreases with larger length. It can be observed that when the responses are too short, the information provided is insufficient to assist the model in generating tweets; conversely, when the responses are too long, they may lead to information redundancy, thereby adversely affecting the model's performance.

We further analyze the impact of the number of responses on the model's performance. As shown in Figure 2(b), the model performs best with only one comment. As the number of responses increases, the model's performance significantly declines. This suggests that introducing multiple human reactions might confuse the model, highlighting the necessity of performing response selection.

Impact of Source Message Length. Subsequently, we analyze the impact of source message length on the model's ability to generate quoting tweets. Figure 2(c) presents the scenario of quoting external links, and a similar trend is also observed in quoting tweets. From the figure, we can observe that when there is minimal source information (0-50), the auto-response augmentation method could help better generate quoting tweets. Moreover, with longer source massages, our model also maintains an improvement in consistency.

Impact of Reward Sample Ratio. Finally, we analyzed the impact of different sample ratios k on the model's performance. As observed in Figure 2(d), the optimal ratios for quoting links and tweets are 0.6 and 0.8, respectively. It is also evident that under all sample ratios, the model's performance surpasses that of LLaMA2. When the sample ratio is 1 (i.e., all samples participate in PPO training), the model's performance has decreased. This indicates that our designed data sampling method significantly aids in generating quoting tweets.

Performance on Different Social Behaviors. In Table 3, we analyze the performance of our model across different social behaviors in two datasets. We divided the test data into three groups based on different popularity factors: like, reply, and retweet. From Table 3, it is evident that our model outperforms LLaMA2 across various social behaviors and exhibits consistency in its performance.

Models	QuoteLink			QuoteTweet		
	Like	**Reply**	**Retweet**	**Like**	**Reply**	**Retweet**
LLaMA2	14.38	14.67	14.89	8.79	8.51	8.34
RaPALM	**16.39**	**16.47**	**16.25**	**12.37**	**12.01**	**11.70**

Table 5: Blue score on different social behaviors.

6.5 Case Study

We conduct a case study in Table 6. It is noticeable that the tweets generated by our model are more detailed and include a greater number of opinions (highlighted by the colored text), which tends to make these tweets more popular among the audience. Furthermore, we can see the impact of auto-responses augmentation. In the response, the viewpoints "have the potential to revolutionize" and "pick up and learn from it" are mentioned. Our

Source Messgae: ChatGPT-A Silver Bullet for Your Customer Support Org? Language models like ChatGPT can write blog posts, hold conversations, and even pass the bar.
Auto Response: I couldn't agree more! Just learn that ChatGPT and other language models ***have the potential to revolutionize*** the customer support industry complex issues. ***Pick up and learn from it***.
LLaMA2: Pondering the future of #customersupport: Will #ChatGPT be the silver bullet for orgs? #AI #languagemodel
RaPALM: Just set up my ChatGPT and ***I'm blown away by its capabilities***! ***Just learn and try it on*** your customer support team. Will it replace human agents? Maybe not, but it's ***definitely a game-changer for customer service***. #ChatGPT

Table 6: Case study of generated response and different models' output for the source message. The text marked with colors denotes opinions, with similar ones indicated by the same color.

model takes these viewpoints into consideration during generation.

6.6 In-depth Analysis of Wording

To further analyze the quality of generated quoting tweets, we conducted a wording analysis. Inspired by Tang (Tan et al., 2014), we analyze wording from four aspects: informativeness, generality, readability, and source message imitation.

Informativeness. We analyzed the contained information from seven perspectives, including sentence length, five parts of speech, and the use of positive and negative words (measured by Connotation Lexicon (Feng et al., 2013)). From the table, we can observe that the more information a tweet contains, the more likely it is to receive dissemination. Additionally, tweets with more positive words and fewer negative words are also helpful. In comparison to the LLaMA2, RaPALM performs better on all indicators related to information.

Generality. According to (Tan et al., 2014), tweets with higher generality are more likely to be disseminated. Following their settings, we used indefinite articles and definite articles to assess generality. For a fair comparison, we set the maximum text length to 30. It can be observed that with the text length being the same, our model includes more indefinite articles and definite articles, enhancing the generality of the tweets.

Readability. We measure readability by using Flesch reading ease (Flesch, 1948) and Flesch-Kincaid grade level (Kincaid et al., 1975). We can conclude, by comparing unpopular and popular tweets, that tweets with higher readability are more likely to be disseminated. Our generated tweets are relatively more readable compared to those generated by LLaMA2. However, they are far less readable than popular tweets, possibly because the generation model uses a richer vocabulary.

	Neg	Pos	LLaMA2	RaPALM
Infomation.				
Length	26.08	30.70	44.24	49.52
Verb	3.80	4.48	5.95	8.31
Noun	8.05	9.57	13.90	16.13
Adjective	1.82	2.14	3.23	4.15
Hashtag	1.04	1.16	3.10	1.76
Positive	1.42	1.63	2.90	3.68
Negative	1.06	1.33	1.91	1.66
Generality.				
Indef	0.54	0.67	0.89	1.30
Def	1.13	1.27	1.78	2.27
Readability.				
Flesch Score ↑	44.71	48.75	23.07	24.71
Flesch Level ↓	13.79	12.12	18.75	14.84
Imitation.				
Unigram	4.03	5.37	24.71	7.33
Bigram	1.73	2.62	18.75	2.91

Table 7: Result of different wording of negative (Neg) and positive tweets (Pos), tweets generated by LLaMa2 and RaPALM. We compare them in four aspects: 1) Informativeness: the information contained in the sentences. 2) Generality: whether the tweet is more general. Indef means indefinite articles (a, an), and Def means definite article (the). 3) Readability: Both a higher Flesch Score and a lower Level indicate easier readability. 4) Imitation: whether it imitates source message.

Imitate Source Massage. Finally, we analyzed whether popular tweets were modeling source messages. Upon comparing two types of tweets (Negative and Positive), we found that neither of them contained a high similarity with news content. This indicates that tweets are not purely narrating the source itself. In contrast, LLaMA2 exhibits a higher degree of news-related information. By including auto-responses, which often contain opinion-based information, our generated quoting tweets exhibit a much lower resemblance.

7 Conclusion

In conclusion, we present the first study on popular quoting tweet generation with two extensive datasets. We propose a novel Response-augmented Popularity-Aligned Language Model (RaPALM) to align language generation with popularity. The experiments show that RaPALM outperforms advanced LLM in generating popular quoting tweets.

Ethics Statement

In our paper, we create a large Twitter dataset for studying popular quoting tweets. We carefully followed Twitter's API guidelines to collect only public tweets and users. The data, used solely for academic research, has been anonymized to protect user privacy, including removing authors' names and replacing specific tags like @mentions and URLs. Adhering to Twitter's redistribution policy, we will only share this anonymized data and require researchers to agree to use it only for academic purposes, ensuring compliance with ethical standards and Twitter's data policies.

Limitations

We list the limitations of our paper in three aspects: 1) Untrained auto-response, 2) lack of author perspective, and 3) generalization of the method.

Untrained auto-response. We understand that people often react to specific details or key information in tweets. Our auto-response generation method directly utilizes the pre-trained language model LLaMA2 without additional training. Consequently, the generated responses tend to be general, lacking in-depth understanding, and targeted responses to specific topics or details. At times, such responses fail to provide a genuine human reaction.

Lack of author perspective. In generating quoting tweets, we considered the reader's perspective by introducing human responses. However, we overlooked the writer's perspective, such as the personal linguistic habits of users when tweeting. As mentioned in (Tan et al., 2014), there is a strong connection between the popularity of a user's tweets and their personal wording.

Generalization of the method. Our RaPALM approach has been validated as effective in quoting tweet generation. In future work, we aim to generalize this approach to different tasks on social media. Because we know that social media texts are short, and many tasks are related to popularity. These are precisely the two directions that our method can address.

In future studies, we will continue to explore quoting tweet generation and expand our RaPALM to different social media tasks.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. RAFT: reward ranked finetuning for generative foundation model alignment. *CoRR*, abs/2304.06767.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784, Sofia, Bulgaria. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Ryuji Kano, Yasuhide Miura, Motoki Taniguchi, Yan-Ying Chen, Francine Chen, and Tomoko Ohkuma. 2018. Harnessing popularity in social media for extractive summarization of online conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1145, Brussels, Belgium. Association for Computational Linguistics.

Yashal Shakti Kanungo, Sumit Negi, and Aruna Rajan. 2021. Ad headline generation using self-critical masked language model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 263–271, Online. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Sotiris Lamprinidis, Daniel Hardt, and Dirk Hovy. 2018. Predicting news headline popularity with syntactic and semantic knowledge using multi-task learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 659–664, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. 2023. How can recommender systems benefit from large language models: A survey.

Yang Liu, Huanqin Ping, Dong Zhang, Qingying Sun, Shoushan Li, and Guodong Zhou. 2023. Comment-aware multi-modal heterogeneous pre-training for humor detection in short-form videos. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 1568–1575. IOS Press.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Yuzhen Niu, Shanshan Chen, Bingrui Song, Zhixian Chen, and Wenxi Liu. 2023. Comment-guided semantics-aware image aesthetics assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1487–1492.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022a. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *NeurIPS*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022c. Training language models to follow instructions with human feedback. In *NeurIPS*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating efficiently extending transformers for long input summarization.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland. Association for Computational Linguistics.

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.

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

Chunpu Xu and Jing Li. 2022. Borrowing human senses: Comment-aware self-training for social media multimodal classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5644–5656, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Structure learning for headline generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9555–9562. AAAI Press.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.