RePALM: Popular Quote Tweet Generation via Auto-Response Augmentation

Anonymous ACL submission

Abstract

A quote tweet allows users to share others' content and add their comments. To help users write a quote tweet with better public engagement, we study the *popular quote tweet generation* task. It aims to generate quote tweets with higher popularity reflected by more likes, replies, and retweets. Despite the exceptional language generation capabilities of large language models (LLMs), limited work has examined how LLMs can learn the popularity of text to engage the public better. Consequently, we propose a novel **Re**sponse-augmented **P**opularity-**A**ligned **L**anguage **M**odel (RePALM) to align language generation to popularity by leveraging readers' insights from augmented auto-responses. Here, we employ the Proximal Policy Optimization framework with a dual-reward mechanism to jointly explore the quote tweet's popularity and consistency with the auto-responses. For experiments, we gathered two datasets of quote tweets with external links and others' tweets. Extensive results show the superiority of RePALM over advanced language models without response augmentation.[1]

1 Introduction

A *quote tweet* is a tweet that shares a source message, e.g., an external link or another user's tweet, by quoting them and adding accompanying comments. Its purpose is to enhance the visibility of the source message, which is beneficial to various applications, such as media broadcasts, advertisements, and social media engagement (Lin et al., 2023). A popular quote tweet can prompt public readers to engage in discussions actively. It essentially helps broaden the dissemination of the source message, inciting 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 quote tweets. To help them better engage the public for meaningful interactions, we present a novel task of **popular quote tweet generation** to extensively study how NLP models can learn to generate a popular quote tweet given a source message of an external link or other users' tweets.

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. However, limited work concerns the public readers' reactions to the text. For this reason, existing models cannot effectively understand the text's popularity, which reflects its potential to draw public engagement. To illustrate this point, Table 1 shows a sample source message of news followed by the manually written and automatic quote

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 Quote 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 #language-model
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
RePALM: ***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 ChatGPT for Customer Service and a manually-written popular quote tweet on the top. Below are three quote tweets generated by different LLMs and our proposed RePALM. The same colors, purple and red, indicate similar meanings.

[1]Our code and datasets are available at https://anonymous.4open.science/r/RePALM-085A/.

tweets. We observe that the manually written reference is rich in original thoughts and opinions. On the contrary, LLaMA2-chat (Touvron et al., 2023) and ChatGPT (Ouyang et al., 2022b) summarize the news without incorporating any additional insights, thus less likely to draw engagement.

Viewing LLMs' limitation in popularity learning, we propose a novel **Re**sponse-augmented **P**opularity-**A**ligned **L**anguage **M**odel (RePALM). RePALM learns to generate popular quote tweets by employing LLMs to predict possible reader responses, which work as a mirror to reflect public reactions for potential engagement measurements. Augmented by these (auto-)responses, RePALM is trained to align the quote 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 RePALM to generate multiple quote tweets. Next, we optimize RePALM's training process with the Proximal Policy Optimization (PPO) framework (Schulman et al., 2017) with a novel dual-reward design. Here, one reward is to predict popularity trained with a popular-unpopular quote tweet pairs; the other measures consistency between generated quote tweets and selected 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, *RePALM is the first model to utilize LLM-predicted auto-responses for popularity-aligned language generation*. By learning from these potential responses, RePALM can effectively generate popular quote tweets that help draw public engagement. For example, as illustrated in Table 1, the output of RePALM is rich in captivating viewpoints, such as "*blown away by its capabilities*" and "*just learn and try it*."

As a pilot study on popular quote tweet generation, we benchmark the task with two datasets: **QuoteLink** with tweets that quote external links and **QuoteTweet** with tweets that quote other users' tweets. There are 70K pairs of popular-unpopular samples; each pair quotes the same source and is from the same author, yet one is more popular.

We further experiment with the two datasets. The main results first show that RePALM outperforms all comparison models in both automatic measure and human evaluation. For example, RePALM achieves 23.26 Rouge-1, compared to 20.94 from ChatGLM3. Besides, the ablation study implies the positive contributions of varying RePALM modules. Then, quantitative analyses show the effectiveness of RePALM in varying scenarios. After that, we conduct a case study to interpret why RePALM can perform better. Next, we analyze the wording of quote tweets from four aspects to examine the essential features of popularity and how RePALM effectively captures them. At last, a case study interprets RePALM's superiority. In summary, our contributions are threefold:

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

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

• We extensively experiment with popular quote tweet generation and show RePALM's superiority.

2 Related Work

Quote Tweet Generation. Although newly proposed, our task can benefit from two lines of methods: summarization and headline generation. The former (Phang et al., 2022; Lewis et al., 2020) aims to extract the salient contents from the source text, and the latter (Kanungo et al., 2021; Zhang et al., 2020a) to create a headline to summarize or quote the source. However, most methods focused on the writing without considering the popularity factors for further public engagements on social media.

Our work is further 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 RePALM explores aligning the language model with popularity for quote tweet generation, which has not been studied previously.

Popularity Analysis. Our task is also 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

Datasets	Pair Number			Token Number			Popularity Gap			Opinion	
	Train	Valid	Test	Src	Pop	UnPop	Like	Reply	Retweet	Pop	UnPop
QuoteLink	18,969	6,323	6,323	186.7	135.1	158.6	299.4	14.1	53.7	3.12	1.57
QuoteTweet	21,892	7,298	7,298	156.1	92.9	118.9	158.1	15.5	57.3	2.97	2.01

Table 2: Statistics of two quote tweets datasets. The Popularity Gap: the average difference in social behaviors, i.e., "Like," "Retweet," and "Reply." For instance, a "Like" value of 299.4 indicates that, on average, Tweet A receives 299.4 more likes than Tweet B. Opinion scores quantify the degree (5-point likert scale) of opinion expression evaluated by GPT-4, and the average is shown. For the GPT-4 evaluation details, we refer readers to Appendix B.

	Emotion	Generality	Readability	Imitation
Popular	1.63	0.67	48.75	5.37
Unpopular	1.42	0.54	44.71	4.03

Table 3: Wording differences between the first 100 tokens of popular and unpopular quote tweets from four perspectives: emotion (positive words), generality (indefinite articles), readability (Flesch reading ease), and degree of imitation of the source message (unigram). See the Appendix C for the detailed evaluation metrics.

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, and we will extensively explore that.

Response Augmentation. Our method is inspired by previous work enriching context with augmented responses to provide readers' views and enhance NLP training. 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, 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.

3 Quote Tweet Datasets

We collected large-scale data from Twitter for our popular quote tweet generation task. Based on the source message types, we separated the data into two distinct datasets: ***QuoteLink*** and ***QuoteTweet***, where the former gathering quote tweets for external links and the latter for other users' tweets.

Data Collection. Following Nguyen et al. (2020), we first downloaded the general Twitter streams from 02/2016 to 10/2018. Then, we removed duplicate users and shortlisted the tweets from users with over 10,000 followers; the reason for that is to investigate tweets with a specific degree of visibility to measure popularity impartially. 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 that quote 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. More data collection details are described in Appendix A.

Tweet Pair Construction. To train models with the popularity of quote tweets, we construct popular-unpopular quote tweet pairs labeled Tweet A and Tweet B to train models with the popularity of quote tweets. We implemented four rules to construct such pairs: 1) Tweets A and B must be from the same author and quote the same source message. 2) Suggested by Tan et al. (2014), 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) To ensure that Tweet A and B have sufficient distinctiveness for learning popularity, we used SimCSE (Gao et al., 2021) to measure the semantic similarity of the tweet pair and removed pairs whose similarity was above the median (0.53 in our datasets). For model training and testing, we randomly split each of the datasets into training (60%), validation (20%), and test (20%) sets.

Data Analysis. Table 2 shows the statistics of two datasets. We observe that in the QuoteLink dataset, the average length of tweets is generally longer than in the QuoteTweet dataset. It indicates

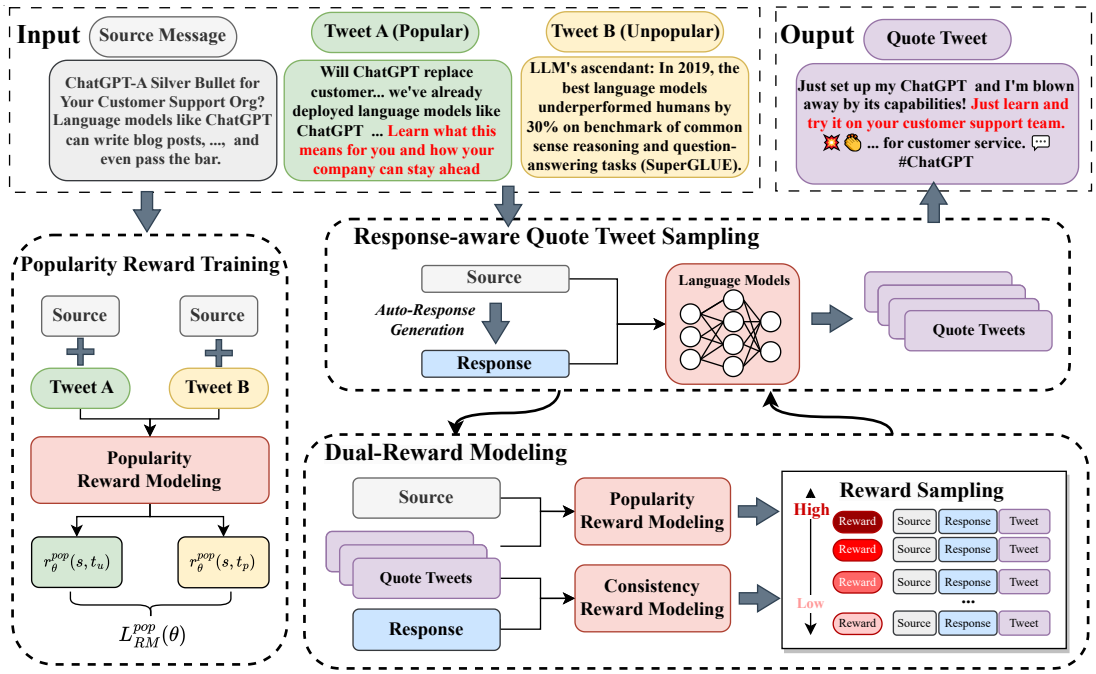


Figure 1: The workflow of RePALM is outlined as follows: the first step involves **generating potential public responses** (§4.1) based on source massages and selecting them based on semantic consistency to the source to yield the auto-response. In the second step, we **generate possible quote tweets with the augmented auto-response** (§4.2) Next, the designed **dual-reward modeling** (§4.3) method aligns the generated quote tweets to popularity. Finally, the training data is chosen for PPO optimization through the **data sampling method** (§4.4).

that users may add more words and detailed information when quoting external links. For the popularity gap, popular quote tweets in both datasets have significantly higher likes, replies, and retweets than unpopular ones. It shows the datasets allow a meaningful comparison of popular-unpopular samples. Moreover, inspired by Maas et al. (2011), we are interested in analyzing how quote tweets' popularity is related to opinions and utilize GPT-4 for opinion assessment. The results show that popular quote tweets contain more opinions than unpopular ones. It highlights the possible benefits of leveraging responses that reflect public opinion for learning popularity (potential engagement).

In addition, Table 3 displays the wording differences between two datasets' popular and unpopular quote tweets. We can observe that popular quote tweets usually exhibit more positive words, more indefinite articles, and higher readability. Popular quote tweets have a slightly higher imitation score, indicating that writing more faithfully to the source message might be more popular. In §6.4, we will discuss how models learn these wording features.

4 RePALM Framework

RePALM overview. To begin with, we describe our datasets as $D = \{s^i, t_u^i, t_p^i\}_{i=1}^N$; s^i stands for the source message, which could be either an external link or a tweet to quote. t_u^i and t_p^i form a pair of unpopular (negative) and popular (positive) tweets of the same quote s^i for the model to compare, and N is the pair number. In the following, we omit the index i for better illustration. RePALM aims to generate a popular quote tweet t_p based on the source s. Its workflow is depicted in Figure 1 with four major components described as follows.

4.1 Auto-Response Generation and Selection

Considering the high relevance of popularity to readers' senses, we incorporate the possible user responses into RePALM's popularity learning process to provide readers' views. However, when quote tweets are created, the public reactions have not yet formed, rendering the absence of actual user responses to refer to. To address this issue, we simulate potential public reactions with an LLM to help RePALM generate popular quote tweets.

Concretely, we first prompt the LLM and employ *top-p sampling* (Holtzman et al., 2020) to sample varying responses (to form the readers' view from diverse angles). Then, we compute their semantic similarity to measure the responses' consistency to the source message. Lastly, we rank the responses based on similarity to select the most relevant ones as the auto-response for the remaining learning

process. The prompt is shown in Appendix D.

4.2 Response-Aware Quote Tweet Sampling

After obtaining the auto-responses, we incorporate them into the quote tweet generation process with the following steps. First, we prompt the LLM to generate quote tweets by augmenting the generated response with the source message. The prompt we adopted is: "*Given the news [source] and potential public reaction [human response], create a quote tweet that highlights the main point of the news while capturing the public's response*." Then, we use *top-p* to sample multiple quote tweets. The purpose is to diversify quote tweets to create a more comprehensive range of samples to train the reward model better (see §4.3) and improve generalization.

4.3 Dual-Reward Modeling

The next step is to align LLMs (with preliminary language generation capabilities) to the popularity factor. Inspired by RLHF (Ouyang et al., 2022a), we exploit the PPO framework and propose dual-reward modeling for popularity alignment. The dual-reward model consists of popularity reward modeling and consistency reward modeling.

Popularity reward modeling primarily assesses how likely social media users will engage with the generated tweet. It is trained on our datasets of comparisons between quote tweets of different popularity for the same source message. Specifically, in the training phase, it takes the source message and two quote tweets as input, i.e., the popular and unpopular ones, and outputs the reward scalar for each quote tweet. We use a cross-entropy loss to optimize the popularity reward model, comparing popularity to labels. The reward difference indicates that one quote tweet will be more popular than the other. The loss function is as follows:

$$\mathcal{L}_{RM}^{pop}(\theta) = -E_{(s,t_u,t_p)\sim D} [\log(\sigma(r_\theta^{pop}(s,t_u) - r_\theta^{pop}(s,t_p)))] \quad (1)$$

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 quote tweet t.

Consistency reward modeling examines the consistency of the generated quote tweet to the potential response. Our intuition is that quote tweets reflecting the readers' viewpoints are more likely to be popular. To achieve this, we measure the semantic similarity between the auto-response and the quote tweet with unsupervised SimCSE as the auxiliary reward. The corresponding loss is $r^{cons}(s,t)$.

The overall reward of RePALM, denoted as $r(s,t)$, is hence the sum of the two rewards:

$$r(s,t) = r_\theta^{pop}(s,t) + r^{cons}(resp,t) \quad (2)$$

4.4 Training Data Sampling and Learning

In PPO-based popularity alignment, the training data quality is crucial, yet social media data can inevitably be noisy. Thus, inspired by Dong et al. (2023), we sample the data to shortlist those with higher model confidence (showing higher reward) for fine-tuning. Specifically, we rank the collected pairs of reward-source-tweet (r,s,t) and select the top k percent of samples with the highest rewards as our sampled training datasets D_{RL}. After that, we adopt the PPO training function defined as:

$$\mathcal{L}_{RL} = -E_{(r,s,t)\sim D_{RL}} r(s,t) \quad (3)$$

5 Experimental Setup

5.1 Model Settings

We will introduce our RePALM model parameters in four parts: 1) **Auto-response generation.** We adopted LLaMA2 (Touvron et al., 2023) (specifically LLaMA2-chat-7b) across all experiments to generate auto-responses. This model is solely for this purpose (without involvement in the quote tweet generation). To sample diverse responses, we set the *top-p* to 0.7 and the temperature to 0.95. For each source message, we sample 5 responses and rank 1 by similarity to the source; 2) **Quote tweet generation.** Another LLaMA2 is employed for generating quote tweets. To sample diverse quote tweets for RL, we maintain the same settings as 1), i.e., *top-p* at 0.7, the temperature at 0.95, and set the sampling number m to 5. The maximum token generation length is set to 150. 3) **Popularity reward modeling.** We used a smaller-scale reward model, GPT-2 (Radford et al., 2019) with a learning rate of $2e^{-4}$, a batch size of 16, and a total of 5 training epochs; 4) **PPO training process.** For PPO, we set the learning rate to $2e^{-5}$, batch size to 4, and training epochs to 3. We set k to 60, i.e., select samples with rewards in the top 60% for training. LoRA (Hu et al., 2022) was used to optimize the quote tweet generation model efficiently.

For training and test, we examine the overall popularity with the sum of likes, replies, and retweets.

5.2 Baselines and Comparison

Recall in §2; we related our task to summarization, headline generation models, and open-source

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
ChatGLM3	20.94	15.49	15.46	0.69	84.11	11.91	8.84	9.21	0.39	82.32
LLaMA2	19.61	14.18	14.57	0.66	83.55	11.59	8.52	8.66	0.37	81.27
LLaMA2-Response	17.21	11.81	12.30	0.56	83.12	11.37	8.03	8.46	0.37	80.43
LLaMA2-FT	18.37	12.13	13.11	0.61	82.01	11.21	8.11	8.27	0.31	80.81
RePALM	**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 4: Main comparison results and ablation results on QuoteLink and QuoteTweet. We report the automatic evaluation metrics R-1 (Rouge-1), R-L (Rouge-L), BLEU, NIST, and BertScore (BertS). Our RePALM 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).

LLM. Our baselines were chosen accordingly. 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, denoted as 3) T5-HeadLine. For open-source LLMs, we included 4) ChatGLM3-6B (Du et al., 2022) and 5) LLaMA2 (Touvron et al., 2023). For comprehensiveness concerns, the comparison also involved our response generation module, 6) LLaMA2-Response and the fine-tuned the LLaMA2 on our datasets, 7) LLaMA2-FT. The details for baseline models are in Appendix E.

5.3 Evaluation Metrics

For ***Automatic Evaluation***, we compare model outputs with popular quote tweets (as references) and evaluate the similarity with ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), NIST (Lin and Hovy, 2003) and BertScore (Zhang et al., 2020b).

For ***Human Evaluations***, we randomly sampled 100 source messages from each dataset, along with quote tweets generated by different models. We then invited 5 human raters to conduct pair-wise comparisons to select the preference between the different quote tweets considering two dimensions: *consistency* of a generated quote tweet to the source message, and *popularity* of the tweet that has the potential to engage the public. This way, we enable easier human ratings to avoid biases. The specific questions for human raters are in Appendix F.

For ***LLM Evaluations***, we used GPT-4 to rate the generated quote tweets on a 5-point Likert scale based on *opinion* and *popularity*. Here, *opinion* measures the expression of a novel viewpoint compared to the source message, considering its crucial roles in popularity (see Table 2). These two criteria are relatively subjective (unlike consistency in human evaluation), and LLM evaluation focuses on them for a more extensive and fair comparison. LLM evaluation concerns all test data in both datasets with detailed prompts in Appendix B.

6 Experimental Results

6.1 Automatic Evaluation

Main Result. Table 4 (top) shows the main comparison result. We draw the following observations.

(1) Generating popular tweets to quote a user's tweet is more challenging than quoting an external link, possibly because user tweets are shorter and lack sufficient context (as shown in Table 2); our RePALM can enrich context via response augmentation and shows superiority. (2) Applying summarization or headline generation models yields subpar performance. It suggests that simply echoing key points from the source message without providing new insights might not be enough to ensure popularity. Meanwhile, the results of LLaMA2-Response are unsatisfactory, indicating that a popular quote tweet entails more than just a random response. (3) Zero-shot ChatGLM3 and LLaMA2 show promising results, indicating the potential of LLMs to serve as the backbone for our task. Meanwhile, LLaMA2-FT performed worse than zero-shot LLaMA2, suggesting the benefits of comparing popular and unpopular samples in learning popularity, a relative concept. (4) Our RePALM, built on LLMs, yielded significantly better results than baselines, showing the effectiveness of response augmentation and RL-based popularity alignment.

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

Table 5: 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.

Ablation Study. To investigate the effects of its components further, we conducted an ablation study with response augmentation, dual-reward modeling, and reward sampling. As seen in Table 4 (bottom), all components, in general, contribute positively to the model's performance. Notably, the model's performance declines the most when responses are reduced, indicating the crucial role of response augmentation in popularity learning.

6.2 Human and LLM Evaluation

To further examine whether the output is helpful to humans, we conduct manual pair-wise evaluations to assess consistency and popularity. RePALM is compared to its backbone, LLaMA2 (also the best baseline). Besides, we experiment with the ablation (-w/o response) to examine the effects of responses. The results are shown in Table 5. RePALM's output is preferred over 1.65 times to the comparison models, indicating the effectiveness of response augmentation and RL-based popularity alignment.

Models	QuoteLink		QuoteTweet	
	Opinion	**Popularity**	**Opinion**	**Popularity**
LLaMA2	2.31	1.34	2.21	1.53
ChatGLM3	2.45	1.47	2.33	1.43
RePALM	**2.88**	**2.34**	**2.78**	**2.12**
-w/o resp	2.36	1.56	2.25	1.54

Table 6: The LLM evaluation results of two datasets, which assess the opinion expression and popularity.

We next present the results of the LLM evaluation in Table 6. RePALM outperforms all comparison models in both criteria with the performance gain especially large in popularity. It is possibly because the augmented responses can helpfully incorporate opinions in the output and further increase the potential to draw public engagements.

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

Table 7: We divided the test set by popularity measures (Like, Reply, or Retweet) and reported BLEU scores.

6.3 Quantitative Analysis

We have shown the overall superiority of RePALM. Here, we examine its results in varying scenarios.

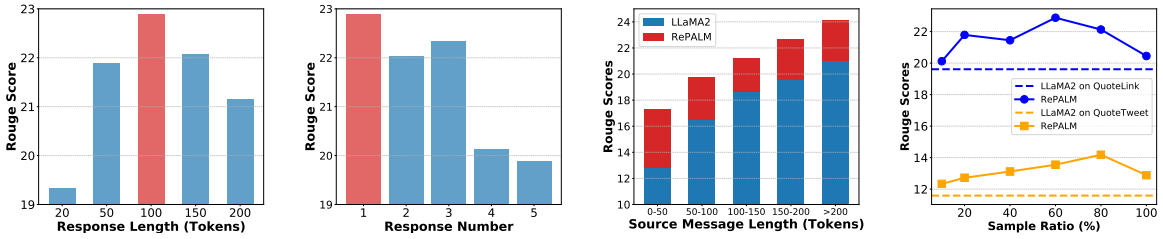
Varying Response Length and Number. While augmented responses shows overall benefits, we quantify their effects here. The first analysis concerns the auto-response length. As shown in Figure 2(a), the score first increases to peak at length 100, then decreases with larger length. It is because augmenting too-short responses offers limited help; conversely, the too-long responses may provide redundancy information and adverse effects.

We then analyze the impact of response numbers on RePALM's performance. Figure 2(b) RePALM the model performs best with only one response. As the number of responses increases, the performance substantially declines. It is possibly because in the current augmentation design, introducing numerous responses might confuse the model, highlighting the usefulness of response selection.

Impact of Source Message Length. We next analyze the impact of source message length. Figure 2(c) shows the results on QuoteLink; a similar trend is observed in QuoteTweet. We observe that when the source messages are very short (0-50 tokens), the augmented auto-responses could help better due to their provision of richer contexts given sparse input. With longer source massages, RePALM also maintains better results in consistency.

Impact of Reward Sample Ratio. Recall that in §4.4, we selected the top k percent data with the highest reward for training. We hence analyzed the impact of different sample ratios k on RePALM's results. Figure 2(d) shows that the optimal ratios for QuoteLink and QuoteTweet are 60% and 80%, respectively. It is also evident that under all sample ratios, RePALM's performance surpasses that of LLaMA2. When the sample ratio is 100% (i.e., all samples participate in PPO training), the model's performance decreases. It indicates that data sampling is helpful in increasing training effectiveness.

Performance on Varying Popularity Measures. The discussions above centered on overall popular-



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

Figure 2: Quantitative analysis results on for hyper-parameters of our model. The first two ((a) and (b)) analyze RePALM since only it introduces the response. In the third and fourth, we incorporate LLaMA2 as the baseline. In (d), the dashed line represents LLaMA2's performance across all data, introduced for easier comparison between RePALM and LLaMA2 across different sample ratios. We report the Rouge-1 score on all experiments

	ChatGLM3	LLaMA2	RePALM	-w/o resp
Emotion.				
Positive	3.17	2.90	**3.68**	3.13
Negative	1.87	1.91	**1.66**	1.77
Generality.				
Indef (a, an)	1.21	0.89	**1.30**	1.13
Def (the)	1.89	1.78	**2.27**	1.35
Readability.				
Flesch Score	22.31	23.07	**24.71**	21.79
Flesch Level	17.22	18.75	**14.84**	19.88
Imitation.				
Unigram	19.87	24.71	**7.33**	17.88
Bigram	14.79	18.75	**2.91**	13.12

Table 8: Wording statistics in the first 100 tokens of quote tweets generated by ChatGLM3, LLaMa2, RePALM, and RePALM(-w/o response). Bold represents wording closer to popular quote tweets, as shown in Appendix 10. As mentioned in Table 3, we evaluate four aspects: emotion, generality, readability, and degree of imitation of the source message. See Appendix C for evaluation specifics.

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
RePALM: 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 9: The case study of generated response and different models' output for the source message. The text marked with colors denotes certain opinions. Similar opinions are indicated by the same color.

ity (the sum of like, reply, and retweet numbers). Here we probe into how RePALM performs on each measure. Table 7 shows the results. The three measures show similar learning difficulties and RePALM consistently outperforms LLaMA2.

6.4 In-depth Analysis of Wording

Recall that in Table 3, we present the differences in wording between popular and unpopular quote tweets in the dataset. In Table 9, we report the performance of different models on these metrics. Similar to popular quote tweets, the quote tweets generated by RePALM outperform in terms of the usage of emotional words, generality, and readability. Notably, the imitation metric dropped from 24.71 to 7.33 compared to LLaMA2, approaching the level of popular quote tweets. It indicates that RePALM avoids copying the source messages while staying faithful to the original text. In particular, RePLAM performs better than its ablation without augmented responses across all metrics. It implies that augmented responses help RePALM generate original contents, helpfully improving popularity.

6.5 Case Study

Finally, a case study in Table 9 interprets why RePALM is effective. The output of RePALM is more detailed and include a richer opinions (highlighted by the colored text), which tends to increase the likelihoods of public engagements. It is because the auto-response contains viewpoints, e.g., "have the potential to revolutionize" and "pick up and learn from it." By response augmentation in popularity alignment, RePALM captures and reflects them in the generation, resulting in better outputs.

7 Conclusion

We have presented the first study on popular quote tweet generation with two extensive datasets. We have proposed a novel model RePALM to exploit augmented auto-responses to better align language generation with popularity. The experiments show RePALM outperforms advanced LLMs in our task.

Ethics Statement

In our paper, we create a large Twitter dataset for studying popular quote 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 quote 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 RePALM approach has been validated as effective in quote 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 quote tweet generation and expand our RePALM to different social media tasks.

References

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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. *Technical report, DTIC Document*.

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.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

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 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 *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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.

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

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.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators.

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. 2020a. 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. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Dataset Construction Detail

First, we downloaded the general Twitter Stream grabbed by the Archive Team [2], containing 400M of Tweet data streamed from 02/2016 to 10/2018 on Twitter. Then, we filter out tweets posted by authors with fewer than 10,000 followers and only keep English tweets that do not contain videos or images. Following that, we have 122,269 users and 259,043 pairs of tweets that report the same source message, which is used to construct positive-negative quote tweet pairs. Finally, after applying our four rules to filter tweet pairs, we obtain the final dataset as shown in Table 2.

B Prompts for LLM evaluation

Recently, using large models as a means of evaluation has become a trend (Zheng et al., 2023; Wang et al., 2023; Chan et al., 2023), achieving higher accuracy than humans in many tasks. Therefore, we utilize GPT-4 (Ouyang et al., 2022b) to rate the opinion and popularity of a quote tweet on a 5-point Likert scale. The prompt for assessing opinion is shown in Figure 7, and the prompt for assessing popularity is shown in Figure 8.

C Wording

In Tables 3 and 8, following Tan et al. (2014), we analyzed the wording differences in the first 100 tokens of various quote tweets, evaluating from four perspectives:, with specific assessment methods outlined as follows: 1) We measure the ***Emotion*** by the number of positive and negative words (measured by Connotation Lexicon Feng et al. (2013)). 2) We use the number of indefinite articles (a,

[2] https://archive.org/details/twitterstream

	Unpopular	Popular
Emotion.		
Positive ↑	1.42	1.63
Negative ↓	1.06	1.33
Generality.		
Indef (a, an) ↑	0.54	0.67
Def (the) ↑	1.13	1.27
Readability.		
Flesch Score ↑	44.71	48.75
Flesch Level ↓	13.79	12.12
Imitation.		
Unigram ↓	4.03	5.37
Bigram ↓	1.73	2.62

Table 10: Complete result of wording analysis of two datasets.

an) and definite articles (the) to assess ***Generality***. 3) For ***Readability***, we use Flesch reading ease (Flesch, 1948) and Flesch-Kincaid grade level (Kincaid et al., 1975). 4) For the evaluation of ***Imitation***, we use the number of shared unigrams and bigrams between the quote tweet and the source message.

In Table 3, for better representation, we selected partial indicators to represent these four perspectives. Positive words serve as the evaluation basis for Emotion, indefinite articles for Generality, Flesch reading ease for Readability, and unigrams for Imitation. The complete result is presented in Table 10.

D Prompt for Auto-response Sampling

Please predict the public's reaction to this source message.
Source message: {source message}

Output:

Figure 3: Prompt for sampling response.

The auto-response sampling and selection process can be formulated as follows:

$$R_{sampled} = LLM(s)$$
$$resp = MaxSim(R_{sampled}, s) \quad (4)$$

where the SimCSE-measured cosine similarity is used to calculate the semantic similarity, which is the same model used in consistency reward modeling. $MaxSim$ function finds the response in $R_{sampled}$ that is most similar to s. Figure D shows the prompt for response generation.

E Prompts for Baseline Models

Please generate a title for this source message.
Source message: {source message}

Output:

Figure 4: Prompt for T5-Headline baseline.

Please generate a quote tweet for this source message.
Source message: {source message}

Output:

Figure 5: Prompt for generating a quote tweet.

We introduced various language models as the baselines and prompted them to generate quote tweets by creating summaries and headlines. In this section, we present the settings for different baselines. For the PEGASUS-X and Bart-Summary models, we directly downloaded models that were fine-tuned, requiring no prompt as the output is the summary of the input. For the T5-Headline, the prompt used is shown in Figure 4. For models that directly generate quote tweets, i.e., T5, LLaMA2, and ChatGLM3, the prompts are shown in Figure 5. For the LLaMA2-Response model, the prompt used is displayed in Figure 3.

F Questions for Human Evaluation

Below is a source message, along with two Quote Tweets generated by different models. Please choose your preference between the models based on two aspects: Consistency and Popularity.
1) Consistency refers to how aligned the Quote Tweet is with the Source Message, ensuring the generated content is relevant to the Quote Tweet.
2) Popularity refers to which Quote Tweet you think has the potential to engage the public.
Source Message: {}
Quote Tweet of Model 1:{}
Quote Tweet of Model 2 :{}

Preference (Choose Model 1 or Model 2):
1) Consistency: 1. Model 1 2. Model 2
2) Popularity: 1. Model 1 2. Model 2

Figure 6: Question Template for human evaluation.

For human evaluation, we conduct a pair-wise comparison. Human raters are asked to select their preference for different quote tweets in the aspect of consistency and popularity. As Shown in Figure 6, the models are anonymous to the human raters.

Predict the popularity of the quote tweet on a scale of 1 to 5 based on its potential to attract attention and engagement on social media. Assign a score according to the following criteria:
1: Predicted popularity is very low, indicating little likelihood of attracting attention from social media users.
2: Predicted popularity is low, suggesting some potential for attention but unlikely to spark widespread discussion or resonance.
3: Predicted popularity is average, indicating some potential for attention but unlikely to become a trending topic.
4: Predicted popularity is high, suggesting significant potential for attention and discussion, likely becoming a trending topic to some extent.
5: Predicted popularity is very high, indicating a strong likelihood of widespread resonance and discussion, making it a trending topic on social media.

Source Message: {}
Quote Tweet:{}

Output:

Figure 7: Prompt for assessing opinion in a quote tweet.

Rate the opinion expressed in the quote tweet on a scale of 1 to 5 based on the clarity, depth, and uniqueness of the viewpoint presented. Assign a score according to the following criteria:
1: The opinion is unclear or absent, lacking in depth or originality.
2: The opinion is present but vague or commonplace, lacking in resonance or discussion potential.
3: There is a moderate degree of personal opinion, although it may lack depth or uniqueness, and may only attract limited attention.
4: The opinion is clear, unique, and likely to provoke resonance or discussion, garnering some degree of recognition.
5: The opinion is distinct, highly individualized, and deeply resonates with the audience, sparking widespread discussion and becoming a trending topic.

Source Message: {}
Quote Tweet:{}

Output:

Figure 8: Prompt for assessing popularity in a quote tweet.