Semantics Preserving Emoji Recommendation with Large Language Models

Zhongyi Qiu
School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, GA, USA
zhongyiqiu@gatech.edu

Kangyi Qiu
School of Computer Science
Georgia Institute of Technology
Atlanta, GA, USA
kqiu37@gatech.edu

Hanjia Lyu
Department of Computer Science
University of Rochester
Rochester, NY, USA
hlyu5@ur.rochester.edu

Wei Xiong
Department of Computer Science
University of Rochester
Rochester, NY, USA
wxiongur@gmail.com

Jiebo Luo
Department of Computer Science
University of Rochester
Rochester, NY, USA
jluo@cs.rochester.edu

Abstract—Emojis have become an integral part of digital communication, enriching text by conveying emotions, tone, and intent. Existing emoji recommendation methods are primarily evaluated based on their ability to match the exact emoji a user chooses in the original text. However, they ignore the essence of users' behavior on social media in that each text can correspond to multiple reasonable emojis. To better assess a model's ability to align with such real-world emoji usage, we propose a new semantics preserving evaluation framework for emoji recommendation, which measures a model's ability to recommend emojis that maintain the semantic consistency with the user's text. To evaluate how well a model preserves semantics, we assess whether the predicted affective state, demographic profile, and attitudinal stance of the user remain unchanged. If these attributes are preserved, we consider the recommended emojis to have maintained the original semantics. The advanced abilities of Large Language Models (LLMs) in understanding and generating nuanced, contextually relevant output make them well-suited for handling the complexities of semantics preserving emoji recommendation. To this end, we construct a comprehensive benchmark to systematically assess the performance of six proprietary and open-source LLMs using different prompting techniques on our task. Our experiments demonstrate that GPT-4o outperforms other LLMs, achieving a semantics preservation score of 79.23%. Additionally, we conduct case studies to analyze model biases in downstream classification tasks and evaluate the diversity of the recommended emojis (https://github.com/VIStA-H/SemanticsPreservingEmojiRec).

Index Terms—semantics preserving, emoji recommendation, large language models, prompt engineering

I. INTRODUCTION

Emojis play a key role in modern digital communication, offering a simple way to convey emotions, tone, and intent that text alone may not fully capture. They are widely used across various platforms, including social media, instant messaging, and even professional digital communications, contributing to more expressive textual experience [1]. However, with the vast number of emojis available, it can be challenging to users to select the most appropriate one to convey their intended
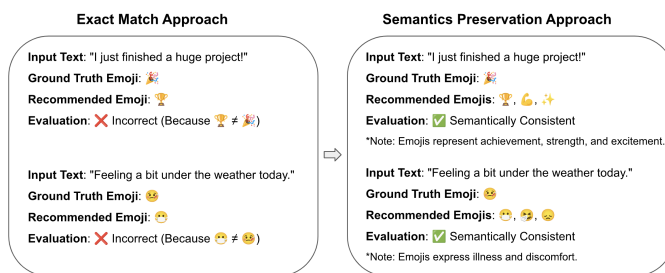
Fig. 1. Comparison of traditional exact match and semantics preservation approaches for emoji recommendation. The semantics preservation approach can suggest multiple emojis that maintain the semantic meaning of the text, even if they differ from the ground truth emojis.

message [2]. To better assist users in selecting appropriate emojis, emoji recommendation tasks have emerged, aiming to provide models with an input text and relevant context to recommend the most contextually appropriate emojis [3].

Existing emoji recommendation methods are primarily evaluated through exact match, where the recommended emoji is compared to the one the user originally selected [4]–[6]. This evaluation criterion, while straightforward, introduces several limitations. First, the same message can be expressed through various emojis, each carrying subtle differences in nuance or connotation [7]. Relying on exact match overlooks this flexibility in how meaning is conveyed through different emoji choices. Second, this evaluation criterion may not accurately reflect the recommendation quality. Existing methods often embed individual user characteristics into the model, as recommendations are influenced by personal emoji usage patterns. Users from different demographic groups frequently use distinct emojis to convey the same sentiment or context [8]–[10]. However, a model's ability to recommend the exact "ground truth" emoji may not fully reflect its understanding of the semantic richness and diversity in emoji use. Fig. 1 provides

intuitive examples of emoji recommendations. Under the traditional exact match criterion, a model would be considered incorrect if it recommends 🏆 instead of 🎉 for "I just finished a huge project!" or 😔 instead of 🤒 for "Feeling a bit under the weather today." However, both 🏆💪✨ and 😵‍💫😷🤢, should be treated as correct since they convey semantically consistent meanings compared with original emojis. Therefore, instead of exact match, a more effective evaluation criterion is required—one that goes beyond merely match the user's chosen emojis and instead captures a model's capability to recommend emojis that align with the meaning of the original text.

To address these issues, we propose a novel evaluation framework for emoji recommendation: *semantics preservation*. This framework aims to assess how effectively models can recommend emojis that maintain semantic consistency, ensuring that the predicted user's affective state, demographic profile, and attitudinal stance are not altered with recommended emojis. To evaluate this, we leverage five downstream classification tasks: sentiment analysis, emotion classification, stance detection, age prediction, and gender prediction.[1] The closer the classification labels of the posts with recommended emojis match those of the posts with the original emojis across these five tasks, the stronger the model's ability to perform semantics preserving emoji recommendation.

With the advent of GPT-series models [11], there has been a growing trend in using LLMs to explore complex questions related to social networks [12], [13] and digital communication behaviors, such as understanding and using emojis [14]. LLMs are particularly well-suited for emoji recommendation, as they are not restricted to a fixed set of emojis and can generate diverse, context-appropriate recommendations for the same input. However, few studies have specifically examined LLMs' ability to perform emoji recommendation. This paper seeks to address these gaps by systematically and comprehensively evaluating the performance of LLMs, including both proprietary and open-source LLMs on the proposed semantics preserving emoji recommendation task.

The contributions of this paper are threefold:

1) We introduce a novel *semantics preserving* evaluation framework for emoji recommendation, specifically designed to assess the ability of models in maintaining the semantic consistency in emoji recommendations.
2) We develop a comprehensive benchmark and design task-specific metrics to effectively evaluate the semantics preserving performance of various LLMs in the context of emoji recommendation.
3) We propose several advanced prompting techniques to improve the performance of the baseline LLMs. We find that by conditioning on user profile information, the semantic consistency of LLMs' emoji recommendation can be significantly improved.

[1] We choose not to include racial classification as classifying races based on social media posts or emoji usage could lead to unintended bias or reinforce harmful stereotypes, which we aim to avoid.

TABLE I
COMPARISON BETWEEN PREVIOUS WORKS AND OUR PROPOSED SEMANTICS PRESERVING EVALUATION FRAMEWORK FOR EMOJI RECOMMENDATION. **MULTI-EMOJI**: ABILITY TO RECOMMEND MULTIPLE EMOJIS FOR EACH SENTENCE. **OPEN VOCAB.**: WHETHER THE MODEL CAN RECOMMEND EMOJIS FROM AN UNRESTRICTED SET (VOCABULARY). **DIVERSITY**: ABILITY TO RECOMMEND DIFFERENT EMOJIS FOR THE SAME INPUT ACROSS MULTIPLE ATTEMPTS. **CRITERION**: THE EVALUATION METHOD USED TO ASSESS EMOJI RECOMMENDATION PERFORMANCE.

Model	Property			Criterion
	Multi-Emoji	Open Vocab.	Diversity	
DeepMoji [15]	✗	✗	✗	Exact match
Mojitalk [16]	✗	✗	✗	Exact match
mmGRU [4]	✗	✗	✗	Exact match
Emoji Decoded [17]	✗	✗	✗	Exact match
Seq2Emoji [6]	✓	✗	✗	Exact match
CAPER [2]	✓	✗	✗	Exact match
EmojiLM [18]	✓	✓	✗	Exact match
MultiEmo [19]	✗	✗	✗	Category match
Human vs. LMMs [14]	✓	✓	✓	Distribution Similarity
Our work	✓	✓	✓	Semantics Preservation

II. RELATED WORK

A. Emoji Recommendation Using Traditional Methods

Emoji recommendation is an active research area in computational linguistics, focusing on developing models to predict emojis based on text input. Various machine learning models have been explored for these tasks. Wu *et al.* [20] developed a model combining CNN and LSTM architectures to recommend emojis in tweets, focusing on 30 different emojis. Felbo *et al.* [15] introduced DeepMoji, a deep neural network leveraging bi-LSTM models with attention mechanisms, trained on a massive dataset of 1.2 billion tweets containing 64 common emojis. Zhou and Wang [16] proposed a conditional variational autoencoder (CVAE) that enhances the traditional seq2seq model by incorporating additional encoders and networks, showing better performance on a dataset of 650,000 tweets. Research by Xie *et al.* [21] indicated that incorporating conversational context could improve emoji prediction accuracy, particularly in dialogues involving two individuals. Kim *et al.* [22] further advanced this approach by combining machine learning with clustering to recommend contextually appropriate emojis based on conversation topics and emotions. Lee *et al.* [19] proposed MultiEmo, a multi-task framework that considers emotion detection to predict the most relevant emoji. They also introduced new evaluation metrics for emoji category prediction, showing superior performance on the Twitter dataset compared to existing models.

Beyond text-based modeling, some research has increasingly explored the use of multiple modalities. These methods combine images, demographics, time, and location, in conjunction with textual data to predict emojis. Barbieri *et al.* [5] employed a multimodal approach using pictures, text, and emojis in Instagram posts, developing models based on bi-LSTM for text and ResNet for visual data, demonstrating that a combination of modalities outperforms text-only models. Zhao *et al.* [4] proposed a multitask, multimodal gated recurrent unit (mmGRU) model that incorporates text, images,

and user demographics to predict emoji categories and their positions within a text, revealing regional differences in emoji usage and understanding. Additional studies, such as those by Barbieri *et al.* [23], explored the temporal aspects of emoji usage, training models on seasonal data to observe changes in emoji semantics. Zhao *et al.* [2] also utilized contextual information, including user gender, current time and emoji co-occurrence factors, to refine emoji recommendations. These studies primarily emphasize exact emoji matching, which, as previously mentioned, has its limitations. In contrast, our evaluation framework focuses on preserving semantic consistency.

B. Emoji Recommendation Using Large Language Models

With the advent of large language models (LLMs), many studies have begun to explore using LLMs for emoji-related research. Kheiri *et al.* [24] examined various GPT-based strategies, such as prompt engineering, fine-tuning, and embedding classification, for sentiment analysis, demonstrating a significant improvement in predictive performance—over 22% in F1-score compared to state-of-the-art models. Zhou *et al.* [17] assessed ChatGPT's effectiveness in emoji-related tasks, suggesting it could serve as an alternative to human annotators by providing clear explanations and reducing misunderstandings in online communication. Lyu *et al.* [14] investigated GPT-4V's interpretation and usage of emojis, finding that it exhibits less interpretative ambiguity and a broader variety in emoji selection compared to humans, indicating partial alignment with human usage patterns. Peng *et al.* [18] built a parallel corpus for text-to-emoji translation using GPT-3.5 Turbo and develops a specialized sequence-to-sequence model for bidirectional translation between text and emojis. However, these studies primarily focus on the capabilities of specific LLMs like GPT in emoji research. In our work, we *extend this exploration to multiple LLMs* and propose novel metrics to evaluate their performance more comprehensively.

C. Comparison with Prior Work

Compared to previous works, our semantics preserving emoji recommendation task introduces an innovative criterion designed to assess how effectively models recommend emojis that maintain **semantic consistency**. Additionally, our framework permits three key properties of emoji recommendation methods, as outlined in Table I.

First, unlike earlier works that typically recommend only a single emoji per sentence, our new task allows **multi-emoji** recommendation, which more accurately mirrors real-world usage where users often combine multiple emojis to express nuanced sentiments. Second, prior approaches often restrict the emoji vocabulary to a predefined set of commonly used emojis. For instance, mmGRU limits recommendations to 35 frequently used emojis per tweet, and CAPER selects from a list of the 50 most popular emojis. In contrast, our task supports an **open emoji vocabulary**, allowing models to recommend any emoji relevant to the text. Finally, our task allows emoji recommendation methods to suggest **diverse**,

contextually appropriate emojis, reflecting a broader range of expressions that better align with real-world usage.

III. Semantics Preserving Emoji Recommendation

Task Formulation: Given the plain text of the original user post denoted as T and the original emojis associated with this text as E_{ori}, a model is tasked with recommending n emojis, denoted as E_{rec} that maintain the semantic content of the original text, as illustrated in the left part of Fig. 2.

Recommendation Evaluation: For each sentence with the predicted or ground truth emojis, we use a strong pre-trained classifier $H(\cdot)$ to infer labels on five downstream tasks. We assess whether the inferred affective state, demographic profile, and attitudinal stance of the user remain unchanged as shown in the right part of Fig. 2. If these attributes are preserved, we consider the recommended emojis to have maintained the original semantics. Specifically, we first use the classifier $H(\cdot)$ to assign labels to the original sentence combining T with E_{ori} on five downstream tasks: sentiment analysis, emotion classification, stance detection, age prediction, and gender prediction. The classification output for each downstream task d (where $d \in \{\text{sentiment, emotion, stance, age, gender}\}$) is denoted as $H_d(T, E_{\text{ori}})$, which serves as the ground truth. Next, we use the same classifier to assign labels to the modified sentences that combine the original plain text T with the recommended emojis E_{rec}. The classification output for the sentence with the recommended emojis is represented as $H_d(T, E_{\text{rec}})$. For each downstream task d, the semantics preserving capability of the model is evaluated by comparing whether $H_d(T, E_{\text{rec}})$ equals the ground truth $H_d(T, E_{\text{ori}})$.

A. Selection of Downstream Tasks

A critical aspect of evaluating semantics preservation in emoji recommendation is the selection of appropriate downstream tasks on social media [25]. We propose that, for a recommended emoji to achieve semantics preservation, it should maintain two key types of information consistent with the original emojis: emotional content and user demographics.

The first type of information is the emotional content conveyed by the sentence. To capture this aspect, we select three downstream tasks: sentiment analysis, emotion classification, and stance detection. Emojis often serve as indicators of **sentiment analysis** [26], [27], so maintaining the sentiment label after replacing the original emoji with a recommended one is crucial for semantics preservation. For this task, we use sentiment labels such as positive, negative, and neutral to evaluate consistency. **Emotion classification** goes beyond general sentiment by identifying specific emotions expressed in the text. Following Ekman's six basic emotions [28], we classify emotions into anger, disgust, fear, joy, sadness, and surprise. We expect that a recommended emoji should align with the six emotional contexts of the original sentence. **Stance detection** is about identifying the author's position or attitude towards a topic. Emojis can modify or reinforce the stance expressed in a sentence, so it is essential that the recommended emojis preserve the stance conveyed by the
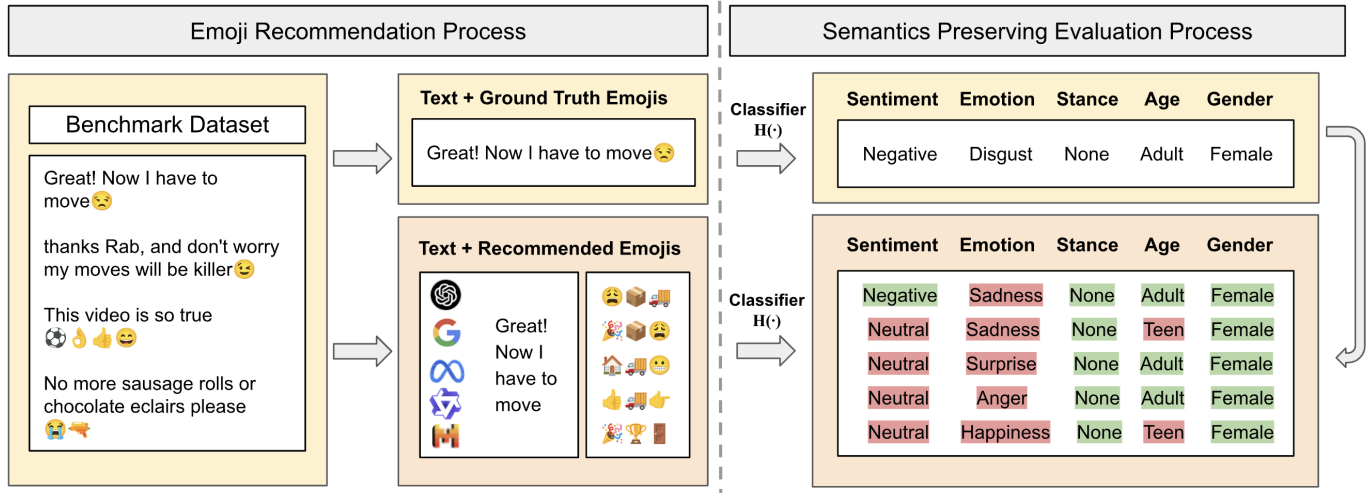
Fig. 2. Overview of the Semantics Preserving Emoji Recommendation Framework. Left side: The Emoji Recommendation Process uses large language models to recommend three emojis for texts from the benchmark dataset. Right side: The Semantics Preserving Evaluation Process compares text + predicted emojis with text + ground truth emojis across 5 selected semantic dimensions, including sentiment, emotion, stance, age, and gender.

original text [29]. We classify stance using the labels none, favor, and against.

The second type of information relates to the user's demographic characteristics. For this part, we select two downstream tasks: age prediction and gender prediction. We opt not to include racial classification as a downstream task to keep the focus of our benchmark on broader aspects of semantics preservation, without introducing sensitive demographic factors. For **age prediction** task, rather than specific ages, we categorize users into broader groups: child, teen, adult, and senior [30]. Different age groups have distinct patterns in emoji usage, even when conveying the same semantic content. [31] Ensuring that the recommended emoji maintains the same age classification as the original helps confirm semantics preservation. For **gender prediction** task, research indicates notable differences in how different genders use emojis to convey the same meaning [32]. The gender labels are classified as male and female. We recognize that gender identity exists on a spectrum, including non-binary, and other gender non-conforming identities, which are frequently underrepresented in existing datasets. However, to maintain the focus on the core objectives of semantics preservation and emoji recommendation, we limit gender prediction to two categories in this study. We encourage the research community to develop and use datasets that reflect the full diversity of gender identities to help mitigate the perpetuation of existing biases.

By evaluating how the predicted emojis align with the original emojis across these five downstream tasks, we can comprehensively evaluate whether the recommended emojis achieve semantics preservation.

B. Downstream Task Classifier

To effectively assess semantics preservation via measuring the performance on downstream tasks, selecting a strong classifier for downstream tasks is essential. Large Language Models have recently emerged as powerful tools, often matching or surpassing human-level classification accuracy across various tasks [25], [33]. In this section, we compare the performance of GPT as a classifier against the traditional top-performant BERTweet model [34] in classification tasks. In each experiment, to mitigate randomness, we query GPT-4o-mini three times and use the majority label as the final prediction. For each query, we set the temperature parameter to 0, while all other hyperparameters as their default values.

1) Accuracy of GPT as a Classifier: We conduct a preliminary experiment focusing on one of the downstream tasks: gender classification. For this experiment, we extract 2,000 tweets with authentic gender labels from the PAN18 dataset [35] for training set and 1,000 tweets for the testing set. BERTweet is fine-tuned on the training set to function as a gender classifier. For the GPT models, both GPT-4o and GPT-4o-mini are prompted to classify gender using the following instruction: *"I will provide you a tweet. Please classify the likely gender of the person who wrote the tweet as 'male' or 'female'. Please only output the answer without justification. Here is the tweet: text"*

Table II presents the gender classification performance of the three models on the test set, including precision, recall, and F1-score. The results demonstrate that both GPT-4o and GPT-4o-mini outperform BERTweet in gender classification.

2) Sensitivity of GPT to emojis: While GPT-4o and GPT-4o-mini achieve a higher accuracy than BERTweet in gender classification, it is important to assess their sensitivity to emojis before adopting GPTs as the classifier for our downstream tasks. To explore this, we design another experiment and assess how emojis affect the GPT models' performance.

In this experiment, we take the original set of tweets used in the gender classification task and replace the existing emojis with three random emojis, selected from the top 500 frequently

used emojis from Unicode[2]. We then prompt GPT-4o and GPT-4o-mini to predict the gender labels for both the original tweets and the tweets with random emojis, using the same instruction format as the first experiment.

To evaluate the consistency of the models' predictions between the two sets of tweets, we calculate the Cohen's Kappa value and correlation between the two predicted label sets. Table III shows that the Kappa value and correlation of GPT-4o-mini is lower than that of GPT-4o, indicating that the predictions of GPT-4o-mini are more affected by the introduction of random emojis. This suggests that GPT-4o-mini is more sensitive to emojis compare to GPT-4o.

In addition to this experimental result, several studies have demonstrated the effectiveness of GPT models as classifiers for sentiment, emotion, stance, age, and gender. For instance, Rathje *et al.* [36] showed that GPT models achieved high accuracy in detecting psychological constructs, such as sentiment and discrete emotions, across multiple languages, often outperforming traditional dictionary-based methods and rivaling top-performing fine-tuned models. Similarly, Liyanage *et al.* [37] evaluated GPT-4 as a stance classifier and found that it performs comparably to human annotators.

Based on the experimental findings and supporting evidence from existing literature, we select GPT-4o-mini as the classifier for our downstream tasks due to its strong performance and cost-effectiveness.

TABLE II
GENDER CLASSIFICATION PERFORMANCE OF BERTWEET, GPT-4O AND GPT-4O-MINI. BEST RESULTS ARE IN BOLD.

Model	Precision	Recall	F1-Score
BERTweet	0.679	0.620	0.633
GPT-4o-mini	0.728	0.737	0.701
GPT-4o	**0.758**	**0.764**	**0.741**

TABLE III
SENSITIVITY OF DIFFERENT GPT VERSIONS TO EMOJIS, WHERE LOWER VALUES INDICATE HIGHER SENSITIVITY TO EMOJIS.

Model	Cohen's Kappa	Correlation
GPT-4o-mini	**0.470**	**0.474**
GPT-4o	0.626	0.627

C. Downstream Task-based vs. Human Judgement-based

After validating the effectiveness of GPT-4o-mini as a classifier, another key aspect that needs to be examined is the effectiveness of the proposed downstream task-based semantics preserving evaluation framework. Accordingly, we compare the semantics preservation scores based on the five downstream tasks with human judgement.

We begin by calculating the number of matching pairs between the original sentence (text + ground truth emojis),

---

[2]Available at: https://home.unicode.org/

and the sentence with emojis generated by GPT-4o (text + recommended emojis). The matching pairs represent the number of labels that are the same across the five downstream tasks, with value ranging from 0 to 5. We use the proportion of correctly matched labels as the semantics preservation score, as shown in Table IV. To ensure a balanced evaluation, we select all data where the number of matching pairs are 0 or 1, and randomly select 50 samples from the data where the number of matching pairs is 2, 3, 4 or 5.

The evaluation dataset is then independently annotated by four researchers. Each researcher performs a binary classification, where a predicted sentence is labeled as 1 if it preserves semantic meaning of the original sentence , and 0 otherwise. The average score for each data sample is then calculated to represent human judgement.

The results, as detailed in Table IV, show that the trends in both the downstream task-based semantics preservation scores and human judgement-based scores are correlated. This alignment further validates our method in evaluating semantics preserving emoji recommendation using downstream tasks.

TABLE IV
COMPARISON BETWEEN THE DOWNSTREAM TASK-BASED AND HUMAN JUDGEMENT-BASED SEMANTICS PRESERVATION SCORES.

Matching Pairs	Count	Downstream task-based	Human judgement-based
0	4	0.0	0.25
1	65	0.2	0.442
2	50	0.4	0.46
3	50	0.6	0.63
4	50	0.8	0.745
5	50	1.0	0.85

IV. BENCHMARK CONSTRUCTION PIPELINE

We construct our benchmark dataset based on the PAN18 dataset [35]. Our goal is to create a high quality dataset that allows us to evaluate the emoji recommendation performance of LLMs on various downstream emoji-related text classification tasks, including sentiment analysis, emotion classification, stance detection, age prediction, and gender prediction.

A. Dataset Filtering

We begin by filtering the English subset of the PAN18 dataset to include only tweets containing one of the top 500 most frequently used emojis, as listed on the Unicode Website[2]. Additionally, duplicate entries are removed, maintaining a diverse set of 20,827 tweets for analysis.

B. Labeling Data for Downstream Tasks

To effectively test the model's ability to recommend emojis across different semantic contexts, we aim for a balanced distribution of labels in the dataset. Therefore, we first employ GPT-4o-mini to classify each of the 20,827 tweets on each of the five downstream tasks. After obtaining the initial labels, we then remove instances where the GPT-4o-mini's predicted label does not align with the predefined class names of the downstream task, ensuring the reliability of the dataset labels.
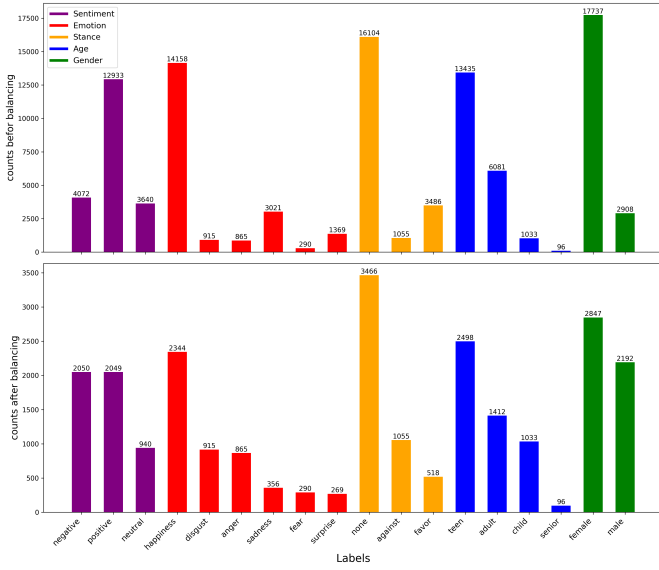
Fig. 3. Comparison of data distribution across five downstream tasks in the benchmark dataset before and after balancing.

C. Dataset Balancing

The distribution of labels after this filtering process is shown in Fig. 3. It is evident that the label distribution in each task is highly imbalanced, particularly in tasks where the number of classes exceeds two, such as emotion classification, age prediction, and stance detection. To address the label imbalance issue, we apply a down-sampling strategy on the filtered data. Specifically, we extract all instances of the minority classes, including:

1) All tweets labeled as *disgust*, *anger*, and *fear* for the emotion classification task.
2) All tweets labeled as *child* and *senior* for the age prediction task.
3) All tweets labeled as *against* for the stance detection task.

Additionally, we randomly select 2,000 instances from the *male* gender class to further balance the gender distribution. After applying these sampling techniques, we compile a benchmark dataset comprising 5,039 tweet posts with an average of 2.25 emojis per sentence. The average sentence length is 14.2 tokens. The final distribution of data across the five tasks is shown in Fig. 3. The resulting benchmark dataset provides a set of examples that are more balanced across all the downstream tasks, enabling a comprehensive evaluation of the emoji recommendation models' ability to maintain semantic integrity across diverse contexts and user demographics.

V. EXPERIMENTS

In this section, we evaluate the performance of existing large language models for the semantics preserving emoji recommendation task on our constructed benchmark dataset. We also provide a detailed analysis on the models' performance across the five downstream tasks and assess the diversity of the emojis used by the models.

A. LLMs as Emoji Recommendation Models

We select the state-of-the-art LLM models as the emoji recommendation methods, focusing on both open-source and proprietary options. For open-source models, we choose LLaMa3.1-70B-Instruct[3], LLaMa3.1-8B-Instruct[4], Qwen2-72B-Instruct [38], Gemma2-9B-Instruct [39], and Mistral-7B-Instruct [40] with parameter sizes ranging from 7 billion to 70 billion, due to their proven performance in various natural language processing tasks. For proprietary models, we selected GPT-4o [11], the most recent model at the time of writing from OpenAI, known for its exceptional performance across various benchmarks.

We also randomly assign three emojis to each sentence as a baseline to compare the performance of model-generated recommendations against random selection in preserving the semantics of the text.

B. Prompting strategies of LLMs for Emoji Recommendation

When using large language models to predict emojis, prompt design plays a critical role in ensuring accurate and contextually relevant results. In this study, we explor three different prompting strategies: 1) Zero-shot, 2) Few-shot, and 3) Conditional Generation. For all strategies, we set the temperature to the default value of 1, which encourages the model to make confident predictions with some level of randomness and creativity in the emoji recommendations. For the open-source models, we access them via the Together AI[4] platform and make API calls to generate predictions.

Zero-shot Prompting In the zero-shot approach, the model is provided with a task-specific instruction without any examples. The prompt contains only instructions explaining what the model should do, and the input is a sentence without any predefined labels or additional guidance. For each user *text*, the prompt inputted to the LLMs is as follows: *"As a social media user, find three emojis that best fit the context: {text}. Do not include any words or explanations—only three emojis."*

Few-shot Prompting Few-shot prompting builds upon the zero-shot approach by providing the model with a few examples before asking it to predict emojis. Each example will demonstrate how to pair text with emojis to help the LLM model better understand the task.

Specifically, we employ a 5-shot approach, carefully selecting diverse examples to cover a wide range of categories across the five tasks. In these five examples, we include texts and emojis representing various emotions, age groups, and genders to ensure each task could learn from this diversity. These examples are sourced from real users and are not part of the benchmark dataset. The five examples are as follows:

1) Context: "After the success of 's walima vid, we're back at it again for 's wedding celebrations!" Emojis: 🎉🎞️🍂

---

[3]Available at: https://www.llama.com/
[4]Available at: https://www.together.ai/

TABLE V
PRECISION, RECALL AND F1 SCORES (ALL IN %) AND DOWNSTREAM TASKS' ACCURACY COMPARISON OF VARIOUS LLMS. AVG DENOTES TO THE
AVERAGE PRECISION ACROSS ALL FIVE TASKS, *i.e.*, THE SEMANTICS PRESERVING SCORE. BEST RESULTS ARE IN BOLD.

LLM	Precision	Recall	F1 score	Sentiment	Emotion	Stance	Age	Gender	Avg
Zero-shot									
GPT-4o	6.93	16.14	9.04	78.41	**67.41**	**88.61**	**82.02**	79.70	**79.23**
LLaMa3.1-70B-Instruct	7.26	**16.89**	**9.52**	78.27	67.18	87.24	81.76	80.39	78.97
Gemma2-9B-Instruct	6.71	15.61	8.70	**78.57**	66.44	87.42	80.59	**81.78**	78.96
Qwen2-72B-Instruct	**7.30**	14.99	9.16	76.80	63.41	87.78	81.52	79.86	77.87
LLaMa3.1-8B-Instruct	5.69	13.17	7.44	76.32	65.71	85.67	78.11	80.49	77.26
Mistral-7B-Instruct	3.70	7.37	4.56	76.68	60.83	86.45	78.43	75.90	75.66
Random Sentence	0.27	0.52	0.33	69.12	57.09	86.96	75.17	74.88	72.65
Few-shot									
GPT-4o	7.39	17.67	9.70	**78.79**	**67.85**	**88.13**	81.68	77.95	78.88
LLaMa3.1-70B-Instruct	**8.17**	**19.50**	**10.86**	78.45	67.43	86.25	82.54	80.43	79.02
Gemma2-9B-Instruct	7.67	17.68	9.97	78.59	66.74	87.97	80.33	**82.36**	**79.20**
Qwen2-72B-Instruct	8.14	16.02	10.04	77.06	63.62	87.93	**82.56**	81.13	78.46
LLaMa3.1-8B-Instruct	6.63	16.11	8.84	77.63	66.62	87.16	80.99	79.96	78.47
Mistral-7B-Instruct	4.21	9.08	5.30	77.40	61.94	88.01	80.67	77.22	77.05
Conditional Recommendation									
GPT-4o	8.23	19.5	10.8	78.61	67.49	88.33	85.69	87.74	81.57
LLaMa3.1-70B-Instruct	**8.56**	**20.27**	**11.34**	78.41	**68.01**	87.70	85.04	87.61	81.35
Gemma2-9B-Instruct	8.31	19.9	10.96	**79.48**	67.73	87.72	**86.58**	**87.89**	**81.88**
Qwen2-72B-Instruct	8.54	17.11	10.58	75.95	62.43	87.64	84.54	87.76	79.66
LLaMa3.1-8B-Instruct	7.08	16.39	9.29	77.16	66.82	86.33	83.39	87.82	80.30
Mistral-7B-Instruct	3.60	8.72	4.69	78.29	62.69	**88.37**	82.73	84.24	79.26

2) Context: ". finally being played in the right position as well. Great performance!" Emojis: 🙌👊⚽

3) Context: "I NEED THE WILSON BLADE 104 IN MY LIFE!!!" Emojis: 😭😭🎾

4) Context: "Just broke me tooth into bits eating a lid of a bottle" Emojis: 😡🙈😡

5) Context: "It's not even 9am and it's already 30 degrees" Emojis: ☀️💧😓

Conditional Emoji Recommendation In the previous zero-shot or few-shot prompting approaches, the input prompt to the LLM is either the original tweet post, or adding predefined examples. These approaches ignore the rich information that could be obtained from the user's profile under certain circumstances. To further improve the recommendation performance, we propose conditional emoji recommendation, where we embed user's gender and age labels as additional context into the prompt to guide the prediction of the emojis. The gender and age labels are predicted by GPT from the original tweet post as a simulation of the real world scenario, where we have access to the demographic information from the user profile. We evaluate whether this conditional recommendation approach can lead to more semantically consistent predictions.

C. GPT-4o-mini as Task Classifier and Label Generator

For evaluating semantics preservation in emoji recommendation, we choose GPT-4o-mini as the classifier for downstream tasks concerning recommendation accuracy, efficiency and cost among a set of LLM models.

For each sentence in the dataset, we feed the text and corresponding emojis to GPT-4o-mini, prompting it to generate labels for each of the five tasks, with a temperature setting of 0. Below is an example of the prompt used for sentiment analysis: *"I will provide you a tweet. Please classify the sentiment of the tweet as positive, negative, or neutral. Please only output the answer without justification. Here is the tweet:{text}"*. The same approach is adapted for emotion, stance, age, and gender classification, each with task-specific classes. These generated labels from the original post will be used as the ground truth label for each post.

For each predicted sentence (original text + predicted emojis), we also use GPT-4o-mini to predict its labels across the five downstream tasks. Then we compare the difference between the predicted labels and ground truth label.

D. Evaluation Metrics

For each user sentence, we use metrics such as F1-score, precision, and recall as the traditional emoji prediction quantitative metrics to evaluate the performance of models in emoji recommendation. These metrics compares the extent of exact match between the model predicted emojis and the ground truth emojis in the sentence. In our work, as described in Section III, we propose a novel evaluation approach, *semantics preserving score*, to measure the semantics preserving capabilities of the models. On each downstream task, the label of the sentence with the predicted emojis are compared with the label of the original sentence, and we calculate the proportion of

Model	Sentence with emojis	Labels on five tasks				
Ground Truth	Roll on the prem at the weekend, international football is dire 🤮	Negative	Disgust	Against	Adult	Male
GPT-4o	Roll on the prem at the weekend, international football is dire ⚽🤮🏆	Negative	Sadness	Against	Adult	Male
Mistral-7B-Instruct	Roll on the prem at the weekend, international football is dire 🏑❤️🏆	Negative	Sadness	None	Teen	Female
Ground Truth	C'mon Dodgers I need to see you guys in the World Series 💪💪💪	Positive	Happiness	Favor	Teen	Male
GPT-4o	C'mon Dodgers I need to see you guys in the World Series ⚾🔵🏆	Positive	Happiness	Favor	Teen	Male
Mistral-7B-Instruct	C'mon Dodgers I need to see you guys in the World Series 🌟🏆❤️	Neutral	Sadness	Favor	Teen	Female
Ground Truth	you should really keep your spurts to yourself 😳	Negative	Disgust	None	Teen	Female
GPT-4o	you should really keep your spurts to yourself 🧑🤐😳	Negative	Disgust	None	Teen	Female
Mistral-7B-Instruct	you should really keep your spurts to yourself 🙅‍♀️🤐😶	Negative	Anger	Against	Teen	Male

correctly predicted cases as the semantics preserving score for this task. The overall semantics preserving score is obtained by averaging the scores across all tasks.

E. Results and Analysis

We analyze six LLM models on this benchmark dataset to evaluate their emoji recommendation and semantics preservation capabilities across the five downstream tasks, as shown in Table V. The results are obtained using three prompting approaches: zero-shot, few-shot, and conditional recommendation. All values are presented as percentages. It is worth noting that, *in this context, if a model's F1 score is not particularly high, but its accuracy on the five downstream tasks is high, this indicates the model's greater capacity for diverse emoji recommendation.* For instance, in the zero-shot setting, GPT-4o achieves the highest average accuracy at 79.23%, although its F1 score of 9.04 ranks only third among the six models. This suggests that GPT-4o excels in recommending diverse emojis that, while differing from the ground truth, maintain the same semantic meaning.

In the few-shot approach, we observe performance improvements in most models except for GPT-4o. The Gemma2-9B-Instruct model demonstrates the best performance. The decline in GPT-4o's performance using the few-shot approach compared to the zero-shot approach could potentially be attributed to GPT-4o's inherently strong language understanding and reasoning capabilities. Providing examples in the prompt might constrain its emoji recommendation flexibility, limiting it to the scope of the five given examples.

In the conditional recommendation approach, we enhance the prompt by including user profile information such as gender and age when predicting emojis. This enhanced prompt significantly improves the performance of all models, with Gemma2-9B-Instruct achieving the highest average accuracy at 81.88%. Across the board, models show an average 2% improvement under this condition, highlighting that providing basic user profile information allows models to generate more accurate emoji recommendations tailored to users' needs.

This benchmark compares the performance of six existing models. We welcome the community to utilize our emoji benchmark to evaluate emoji recommendation and semantics preservation capabilities across a broader range of models.

F. Case Study

In this section, we present two key analyses that provide deeper insights into the performance of large language models on the semantics preserving emoji recommendation task. One type of case study focuses on analyzing the classification accuracy of each specific class in each downstream task. Another case study examines the diversity capability of the models in emoji recommendation.

1) Categorical Bias in Emoji Recommendation: Table VII shows a detailed per-class accuracy on each task. We observe class bias in all the six LLM recommendation models. For example, the accuracy of the female class in gender classification consistently surpasses that of the male class across all models. In the sentiment classification task, the neutral class shows noticeably lower accuracy compared to the positive and negative classes. A similar trend is observed in emotion classification, where models perform much better on the happiness class, while the disgust class exhibits poor performance. In the age classification task, the accuracy of the senior class is significantly lower than other classes, with a value around 20%.

Recent studies [41] reveal that such categorical bias often arises when LLM models, faced with uncertainty in prediction, revert to heuristics and patterns ingrained from public datasets, resulting in predictions that conform to the most prevalent or recognizable trends within the training data. For instance, the bias towards the senior class may stem from the underlying assumption that most users posting on social platforms like Twitter are teenagers or middle-aged adults. Consequently, the models struggle to accurately classify posts from older users, often defaulting to the teen category. Similarly, models tend to assume that users predominantly express happiness in their posts. When uncertain about user's actual emotion, the models default to happiness, resulting in significantly higher classification accuracy for this class compared to other emotional categories.

2) Diversity of Emoji Recommendation: In this section, we evaluates how diverse the set of emojis recommended

LLM	Sentiment			Emotion						Stance			Age				Gender		Avg.(5)
	Positive	Negative	Neutral	Happiness	Sadness	Fear	Anger	Surprise	Disgust	Favor	None	Against	Child	Teen	Adult	Senior	Male	Female	
GPT-4o	87.51	81.37	52.13	84.00	57.30	61.38	67.05	**45.35**	37.60	78.38	91.78	83.22	70.28	91.91	77.27	20.83	70.30	86.93	79.23
GPT-4o$_{few-shot}$	88.43	81.85	51.06	85.71	59.55	59.31	67.86	40.52	36.07	**78.76**	91.63	81.23	66.51	91.63	79.25	21.88	68.61	85.14	78.88
GPT-4o$_{CR}$	91.17	80.88	46.28	87.97	57.30	53.45	63.82	33.09	37.05	77.41	92.04	81.52	79.38	94.20	79.24	27.08	75.00	**97.54**	81.57
LLaMa3.1-70B-Instruct	86.14	84.68	47.13	82.81	53.93	63.79	67.17	40.52	37.60	74.90	89.56	85.69	67.57	92.95	76.49	20.83	70.30	88.16	78.97
LLaMa3.1-70B-Instruct$_{few-shot}$	88.92	83.61	44.36	85.54	52.25	57.59	67.05	32.34	40.77	75.87	90.25	**86.25**	70.18	92.99	77.34	19.76	67.75	90.20	79.20
LLaMa3.1-70B-Instruct$_{CR}$	89.70	83.46	42.87	86.90	52.80	59.65	65.43	27.51	**42.51**	77.99	89.87	85.31	**80.83**	93.95	76.49	23.96	75.13	97.22	81.35
Gemma2-9B-Instruct	86.43	84.00	49.57	82.34	53.93	60.34	71.21	42.01	35.19	76.83	90.19	83.51	66.31	91.67	75.42	21.87	67.70	92.63	78.96
Gemma2-9B-Instruct$_{few-shot}$	86.97	83.80	48.94	83.66	55.06	65.52	70.98	40.52	32.02	77.03	90.91	83.70	64.47	92.03	74.93	26.04	68.11	93.33	79.20
Gemma2-9B-Instruct$_{CR}$	88.73	**85.51**	46.17	85.15	57.87	57.24	**72.83**	37.17	34.43	78.57	90.13	84.27	79.77	**95.44**	78.05	**55.21**	76.64	96.56	**81.88**
LLaMa3.1-8B-Instruct	84.29	81.80	47.02	81.23	56.74	63.45	66.71	31.60	39.23	74.71	89.21	79.43	68.83	84.71	77.20	19.79	71.03	87.78	77.26
LLaMa3.1-8B-Instruct$_{few-shot}$	87.41	82.10	46.60	84.94	54.78	**66.55**	64.28	32.71	36.50	72.59	90.13	84.55	65.05	93.35	74.93	19.79	69.39	88.09	78.47
LLaMa3.1-8B-Instruct$_{CR}$	89.07	81.22	42.34	86.43	60.11	59.31	63.58	26.39	36.50	74.13	89.01	83.51	76.28	92.15	76.49	33.33	**77.24**	95.96	80.30
Qwen2-72B-Instruct	88.97	75.37	53.40	86.56	63.48	58.28	52.14	39.78	23.28	78.38	**93.05**	75.07	69.41	90.99	77.76	20.83	68.29	88.76	77.87
Qwen2-72B-Instruct$_{few-shot}$	88.34	76.00	**54.79**	86.95	67.70	55.86	48.55	42.38	25.25	77.41	92.87	76.87	70.28	91.19	**80.38**	21.88	70.07	89.64	78.46
Qwen2-72B-Instruct$_{CR}$	**91.70**	73.61	46.70	**90.32**	**70.22**	47.59	42.08	33.46	20.44	78.76	92.99	74.41	78.41	92.39	79.46	20.83	76.32	96.56	79.66
Mistral-7B-Instruct	82.82	83.46	48.51	79.44	67.98	61.72	54.10	33.09	24.59	71.43	91.06	78.67	65.05	90.63	70.68	18.75	64.10	85.00	75.66
Mistral-7B-Instruct$_{few-shot}$	84.97	85.41	43.40	81.44	63.76	63.45	53.87	31.60	27.32	71.04	92.27	82.37	67.57	92.99	72.88	15.63	63.73	87.60	77.05
Mistral-7B-Instruct$_{CR}$	87.95	84.44	43.83	83.19	**70.22**	60.00	55.84	31.97	23.61	72.97	92.79	81.42	79.96	92.43	72.17	15.63	67.61	97.05	79.26
Random Sentence	67.84	73.66	62.02	67.19	51.12	52.76	58.73	37.17	39.23	55.98	96.02	72.42	60.02	93.88	57.15	16.66	50.32	93.78	72.65

by different models is across the entire dataset. Table VIII presents the unique emoji count for each model, sorted from the highest to the lowest. Among the models, GPT-4o stands out by recommending the largest number of unique emojis, while LLaMa3.1-8B exhibits the least diversity, with only 534 unique emojis. This variation in the number of unique emojis suggests that GPT-4o can potentially lead to more diverse emoji usage in real-world applications.

TABLE VIII
UNIQUE EMOJI COUNT FOR DIFFERENT MODELS, SORTED FROM HIGH TO LOW.

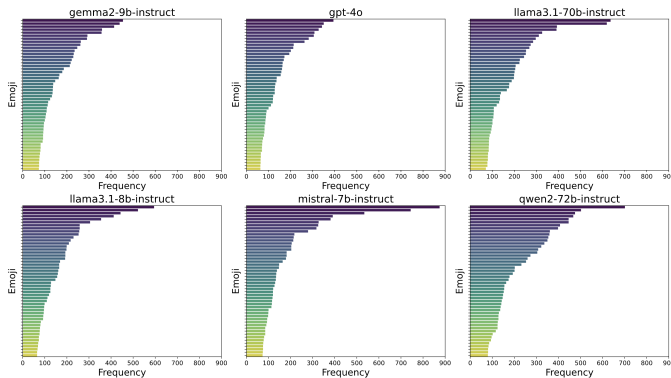| Model	Unique Emoji Count
GPT-4o	**803**
Mistral-7B-Instruct	717
Gemma2-9B-Instruct	664
LLaMa3.1-70B-Instruct	570
Qwen2-72B-Instruct	554
LLaMa3.1-8B-Instruct	534

Fig. 4. Distribution of Top 50 Frequently Used Emojis Across Different Models

Fig. 4 complements this analysis by showing the distribution of the top 50 most frequently used emojis across six models. The distribution patterns reveal that models like Mistral-7B, Llama3.1-8B, Llama3.1-70B and Qwen2-72B have a higher concentration on top 5 emojis, which are used much more frequently compared to others. On the other hand, GPT-4o and Gemma2-9B display a more balanced distribution, suggesting that they recommend a wider range of emojis more equally.

VI. CONCLUSION AND FUTURE WORK

In this paper, we present a novel evaluation framework for semantics preserving emoji recommendation and a comprehensive benchmark assessing the performance of various LLMs on this task. Our benchmark incorporates multiple evaluation metrics and investigates the impact of different prompting methods on model performance. We have demonstrated that Large Language Models, particularly GPT-4o, have significant potential in generating diverse and contextually appropriate emoji recommendations, especially when augmented with demographic information.

While our work provides valuable insights, there are some limitations. First, the focus on binary gender categories—male and female—may overlook non-binary and other gender identities. To address this, future iterations could include an additional, inclusive category alongside male and female, allowing for a broader representation. Second, model performance may be affected by data imbalance, which could introduce bias. Moving forward, we plan to develop a more complex and challenging dataset for large language models, ensuring a more balanced data distribution to improve fairness and robustness.

For future work, we aim to expand benchmark by incorporating non-English corpora and expanding the dataset with more recent and diverse data from various platforms beyond X (formerly Twitter). This will improve the linguistic and cultural diversity of the benchmark. Additionally, the noticeable categorical bias of the existing LLM models analyzed in Section V-F motivates us to explore more accurate and

unbiased emoji recommendation approaches, such as using Retrieval-Augmented Generation (RAG) [42] to obtain more unbiased data or fine-tuning debiased Large Language Models.

REFERENCES

[1] Y. A. Telaumbanua, N. T. N. Telaumbanua, M. D. Halawa, B. Gulo, D. M. E. Halawa, and E. K. Waruwu, "The use of emojis in language communication on social media platforms," *Journal of English Language and Education*, vol. 9, no. 4, pp. 118–130, 2024.

[2] G. Zhao, Z. Liu, Y. Chao, and X. Qian, "Caper: Context-aware personalized emoji recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 9, pp. 3160–3172, 2020.

[3] F. Barbieri, J. Camacho-Collados, F. Ronzano, L. E. Anke, M. Ballesteros, V. Basile, V. Patti, and H. Saggion, "Semeval 2018 task 2: Multilingual emoji prediction," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 24–33.

[4] P. Zhao, J. Jia, Y. An, J. Liang, L. Xie, and J. Luo, "Analyzing and predicting emoji usages in social media," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 327–334.

[5] F. Barbieri, M. Ballesteros, F. Ronzano, and H. Saggion, "Multimodal emoji prediction," *arXiv preprint arXiv:1803.02392*, 2018.

[6] D. Peng and H. Zhao, "Seq2emoji: A hybrid sequence generation model for short text emoji prediction," *Knowledge-Based Systems*, vol. 214, p. 106727, 2021.

[7] S. Wijeratne, L. Balasuriya, A. Sheth, and D. Doran, "A semantics-based measure of emoji similarity," in *Proceedings of the international conference on web intelligence*, 2017, pp. 646–653.

[8] L. L. Jones, L. H. Wurm, G. A. Norville, and K. L. Mullins, "Sex differences in emoji use, familiarity, and valence," *Computers in Human Behavior*, vol. 108, p. 106305, 2020.

[9] Y. Chen, X. Yang, H. Howman, and R. Filik, "Individual differences in emoji comprehension: Gender, age, and culture," *Plos one*, vol. 19, no. 2, p. e0297379, 2024.

[10] Y. Wirza, H. Hanifah, H. Hanifah *et al.*, "The difference in emoji usage between genders," in *Twelfth Conference on Applied Linguistics (CONAPLIN 2019)*. Atlantis Press, 2020, pp. 232–236.

[11] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.

[12] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can large language models transform computational social science? arxiv," *Preprint posted online on April*, vol. 12, 2023.

[13] Y. Zhu, P. Zhang, E.-U. Haq, P. Hui, and G. Tyson, "Can chatgpt reproduce human-generated labels? a study of social computing tasks," *arXiv preprint arXiv:2304.10145*, 2023.

[14] H. Lyu, W. Qi, Z. Wei, and J. Luo, "Human vs. lmms: Exploring the discrepancy in emoji interpretation and usage in digital communication," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, 2024, pp. 2104–2110.

[15] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," *arXiv preprint arXiv:1708.00524*, 2017.

[16] X. Zhou and W. Y. Wang, "Mojitalk: Generating emotional responses at scale," *arXiv preprint arXiv:1711.04090*, 2017.

[17] Y. Zhou, P. Xu, X. Wang, X. Lu, G. Gao, and W. Ai, "Emojis decoded: Leveraging chatgpt for enhanced understanding in social media communications," *arXiv preprint arXiv:2402.01681*, 2024.

[18] L. Peng, Z. Wang, H. Liu, Z. Wang, and J. Shang, "Emojilm: Modeling the new emoji language," *arXiv preprint arXiv:2311.01751*, 2023.

[19] S. Lee, D. Jeong, and E. Park, "Multiemo: Multi-task framework for emoji prediction," *Knowledge-Based Systems*, vol. 242, p. 108437, 2022.

[20] C. Wu, F. Wu, S. Wu, Y. Huang, and X. Xie, "Tweet emoji prediction using hierarchical model with attention," in *Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers*, 2018, pp. 1337–1344.

[21] R. Xie, Z. Liu, R. Yan, and M. Sun, "Neural emoji recommendation in dialogue systems," *arXiv preprint arXiv:1612.04609*, 2016.

[22] J. Kim, T. Gong, B. Kim, J. Park, W. Kim, E. Huang, K. Han, J. Kim, J. Ko, and S.-J. Lee, "No more one liners: Bringing context into emoji recommendations," *ACM Transactions on Social Computing*, vol. 3, no. 2, pp. 1–25, 2020.

[23] F. Barbieri, L. Marujo, P. Karuturi, W. Brendel, and H. Saggion, "Exploring emoji usage and prediction through a temporal variation lens," *arXiv preprint arXiv:1805.00731*, 2018.

[24] K. Kheiri and H. Karimi, "Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning," *arXiv preprint arXiv:2307.10234*, 2023.

[25] H. Lyu, J. Huang, D. Zhang, Y. Yu, X. Mou, J. Pan, Z. Yang, Z. Wei, and J. Luo, "Gpt-4v (ision) as a social media analysis engine," *arXiv preprint arXiv:2311.07547*, 2023.

[26] M. Shiha and S. Ayvaz, "The effects of emoji in sentiment analysis," *Int. J. Comput. Electr. Eng.(IJCEE.)*, vol. 9, no. 1, pp. 360–369, 2017.

[27] T. Chen, W. Xiong, H. Zheng, and J. Luo, "Image sentiment transfer," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4407–4415.

[28] P. Ekman, "Are there basic emotions?" 1992.

[29] E. Zotova, R. Agerri, and G. Rigau, "Semi-automatic generation of multilingual datasets for stance detection in twitter," *Expert Systems with Applications*, vol. 170, p. 114547, 2021.

[30] D. Nguyen, D. Trieschnigg, A. S. Dogruöz, R. Gravel, M. Theune, T. Meder, and F. De Jong, "Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment," in *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*. Association for Computational Linguistics, 2014, pp. 1950–1961.

[31] S. R. Jaeger, Y. Xia, P.-Y. Lee, D. C. Hunter, M. K. Beresford, and G. Ares, "Emoji questionnaires can be used with a range of population segments: Findings relating to age, gender and frequency of emoji/emoticon use," *Food Quality and Preference*, vol. 68, pp. 397–410, 2018.

[32] T. K. Koch, P. Romero, and C. Stachl, "Age and gender in language, emoji, and emoticon usage in instant messages," *Computers in Human Behavior*, vol. 126, p. 106990, 2022.

[33] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46595–46623, 2023.

[34] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," *arXiv preprint arXiv:2005.10200*, 2020.

[35] F. Rangel, P. Rosso, M. Montes-y Gómez, M. Potthast, and B. Stein, "Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter," *Working notes papers of the CLEF*, vol. 192, 2018.

[36] S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjieh, C. E. Robertson, and J. J. Van Bavel, "Gpt is an effective tool for multilingual psychological text analysis," *Proceedings of the National Academy of Sciences*, vol. 121, no. 34, p. e2308950121, 2024.

[37] C. Liyanage, R. Gokani, and V. Mago, "Gpt-4 as a twitter data annotator: Unraveling its performance on a stance classification task," *Authorea Preprints*, 2023.

[38] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[39] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[40] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[41] E. Ferrara, "Should chatgpt be biased? challenges and risks of bias in large language models," *arXiv preprint arXiv:2304.03738*, 2023.

[42] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.