Towards Reliable Evaluation of Emotional Text Generation in LLMs: Human vs. Automatic Metrics

Sadegh Jafari, Els Lefever and Véronique Hoste
LT3, Language and Translation Technology Team
Ghent University, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
{sadegh.jafari, els.lefever, veronique.hoste}@ugent.be

Abstract

Evaluating emotion generation in large language models (LLMs) remains a challenging problem due to the subjective nature of emotions and the lack of reliable automatic evaluation metrics. In this paper, we introduce a robust and extensible benchmark for systematically assessing automatic metrics in emotion generation tasks. The benchmark currently includes 13 automatic evaluation metrics and five state-of-the-art LLMs, and can be easily extended without requiring additional human annotations. Through a correlation analysis with human evaluations on a carefully curated annotated subset, we identify the emotion recognition score (ERS) metric, computed with DeepSeek v3.2 and GPT-5-nano, as the most reliable automatic evaluator, achieving correlations exceeding 0.98 and 0.95. Interestingly, despite relying on the same underlying LLM (GPT-5-nano), the emotion absolute score (EAS) metric shows a negative correlation, demonstrating that LLM strength alone does not guarantee automatic metric alignment with human judgment. We also provide lightweight, non-LLM-based alternatives, suitable for low-resource settings where large models are not accessible. A comprehensive per-class emotion analysis further highlights the strengths and weaknesses of the evaluated models. Overall, our results offer a practical and scalable framework for benchmarking emotion generation evaluation metrics and pave the way for more reliable, fair, and interpretable emotional language evaluation.

Keywords: Emotion Generation, Automatic Evaluation, Benchmarking

1. Introduction

Human emotions shape how we communicate, build trust, and make decisions, and they are therefore central to many real-world interactions (De Melo et al., 2010). In sensitive domains such as healthcare, education, or crisis response, AI systems must do more than identify emotions: they should be capable of generating responses that are emotionally appropriate and trustworthy. Recent work argues for this shift from emotion recognition to affective text generation (Singh et al., 2020). However, a persistent obstacle remains: how to evaluate the emotion generation performance of generated text in a reliable and scalable way. Prior studies and benchmarks have highlighted this gap, showing that automatic metrics and human evaluations often focus on different aspects of emotional competence and can lead to inconsistent conclusions (Chen et al., 2024b; Sabour et al., 2024).

Benchmarks such as EmotionQueen (Chen et al., 2024b) and EmoBench (Sabour et al., 2024) have advanced evaluation by formalizing tasks for emotion recognition, empathetic response, and emotion reasoning, and by combining automatic measures with human judgments. Still, these efforts primarily target models' understanding; less attention has been paid to whether commonly used automatic metrics reliably reflect the emotion recognition and generation performance humans perceive in model outputs.

This paper addresses that gap by turning the evaluation lens onto the evaluation metrics themselves. We prompt five LLMs to produce emotionally conditioned outputs using the UniC dataset (Du et al., 2025) and compute a suite of 13 automatic metrics that mostly (11 metrics) do not require labeled references, offering a scalable alternative to reference-based scoring. To assess metric validity, we collect human ratings and analyze correlations between automatic scores and perceived emotion generation performance.

Unlike most prior work, which centers on improving or benchmarking recognition methods, our focus is methodological: we investigate which automatic metrics are robust indicators of emotion generation performance in LLM-generated texts. By validating metrics against human judgments across different LLMs, we aim to provide practical guidance for future evaluations of affective text generation and to surface which metrics most faithfully capture the human notion of emotional expression in generated language.

2. Related works

We review two complementary lines that inform our study: (i) papers that introduce new models or methods for emotional text generation and report evaluation for those models, and (ii) benchmarking studies that systematically assess generative models in their capabilities of emotional content generation. Our primary interest is how emotion generation has been evaluated in the literature, because our contribution is to examine and validate

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2836–2847
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

the evaluation metrics themselves.

Research that proposes new emotion generation methods tends to fall into three broad categories: full-model fine-tuning, parameter-efficient/adapter methods, and prompt-based techniques. Full-model fine-tuning adapts pretrained LLMs on emotion-rich corpora so the model internalizes affective patterns; for example, Zhang et al. (2025) fine-tunes LLaMA-2-7B on multi-party dialog data with textual and visual emotion cues and reports gains on emotion recognition in conversation (ERC) datasets. Parameter-efficient or adapter-style methods avoid full re-training and prioritize low computational cost. Dong et al. (2025) illustrates this with a latent-space steering approach that extracts and injects emotion vectors per layer at inference time. Finally, prompt-based approaches steer pretrained LLMs without weight updates by augmenting prompts with emotion-specific cues or psychological stimuli; representative works include Jafari et al. (2025), Vinay et al. (2024), or Sahoo et al. (2024).

A second strand of work centers on benchmarking and evaluation. These studies design tasks and metrics to probe different facets of emotional intelligence and generation quality. For example, EmotionQueen (Chen et al., 2024b) evaluates recognition and empathetic response through multiple task types (Key Event Recognition, Implicit Emotion Recognition, etc.); EQ-Bench (Paech, 2023) targets emotional understanding and intensity prediction in dialog settings; and LongEmotion (Liu et al., 2025b) studies emotional intelligence in long-context scenarios across classification, QA, conversation, and summarization tasks. Table 1 summarizes the evaluation scope across representative model papers and benchmarking studies. It reveals an important pattern: not all model papers (the first group) aim for comprehensive or dataset-independent evaluation; many report a focused set of automatic metrics tailored to their method, whereas benchmarking studies generally aim at a broader, systematic evaluation across multiple metrics and tasks.

Table 2 summarizes key aspects of human evaluations conducted in a subset of prior studies that employed user-based assessments. Among these, EmotionQueen and LongEmotion stand out as representative examples of human-centered evaluation, differing in their scale, experimental protocols, and whether judgments were elicited on human- or machine-generated outputs. These evaluation designs offer valuable insights into how to collect consistent and meaningful human judgments. However, their primary aim was to compare or validate model performance rather than to assess the reliability of automatic metrics used for such comparisons. Moreover, EmotionQueen is not publicly available; its repository only provides a few illustrative samples (fewer than ten), and while LongEmotion is accessible, it includes only the textual data without the corresponding ground-truth labels. Consequently, we were unable to incorporate either of these datasets into our evaluation, so we used the UniC (Du et al., 2025) dataset. Our work focuses specifically on the evaluation aspect: we systematically examine which automatic metrics best align with human judgments across diverse datasets and LLMs, seeking to identify robust, dataset-independent metrics for emotion text generation.

3. Dataset

Given the limitations of existing datasets discussed earlier, we employ the UniC dataset (Du et al., 2025), which was curated from YouTube monologues (e.g., book and movie reviews). This dataset captures multimodal, non-acted, implicit, and naturally expressed emotions, making it well-suited for our study. UniC comprises 964 video clips derived from a set of source videos and was produced through a multi-step curation pipeline (keyword search, subtitle filtering, and manual validation). Clips are short (roughly 10 seconds) and were annotated separately for four modalities (text, audio, silent video, and all modalities combined) using both categorical labels (initially 26 categories reduced to seven emotions through clustering) and dimensional scores (valence and arousal). A small subset of the dataset includes multiple independent annotations: two source videos (61 clips) were each annotated by three annotators, while the remaining clips carry a single annotation. In this study, we use only the text-modality annotations (transcripts), since our focus is on evaluating automatic metrics for generated emotional text derived from UniC's natural, implicit expressions.

4. Generation

We conduct prompt-based text generation on the UniC annotations (transcripts) to create emotionally conditioned variants of naturally occurring monologue utterances. For each transcript, we first apply a neutralization prompt to remove explicit emotional expressions while preserving meaning and overall sentence structure. We then apply an emotion re-injection prompt, asking the model to express a target emotion with minimal deviation from the neutralized text. This two-step procedure (neutralize $\rightarrow$ re-inject) allows us to tightly control semantic content and directly compare each utterance across three versions: original, neutralized, and emotionalized.

Paper	# Automatic Metrics	Human Evaluation	Understanding Evaluation	Generation Evaluation	# Samples (Dataset)
1. Models for Emotional Text Generation					
Dong et al. (2025)	4	✗	✗	✓	500
Zhang et al. (2025)	3	✗	✓	✗	5.2k
Jafari et al. (2025)	6	✗	✗	✓	5.8k
2. Benchmarking and Evaluation of Emotional Text Generation					
EmotionQueen (Chen et al., 2024b)	4	✓	✓	✓	10k
EQ-Bench (Paech, 2023)	1	✗	✓	✗	60
LongEmotion (Liu et al., 2025b)	5	✓	✓	✓	1k
Our Work	13	✓	✓	✓	4.8k

Table 1: Comparison of related works on emotional text generation and benchmarking.

Paper	Human/Machine generated	# Annotators	# Samples
EmotionQueen (Chen et al., 2024b)	Machine	3	1k
LongEmotion (Liu et al., 2025b)	Human	8	370
Our Work	Human	175	1.4k

Table 2: Human evaluation details of related works (only for papers that conducted human studies).

Text generation is performed using five LLMs: LLaMA-3.3-70B Instruct FP16 (Grattafiori et al., 2024), Mistral NEMO-12B Instruct FP8 (AI, 2024), GPT-4o-Mini (OpenAI, 2025c), GPT-4.1 (OpenAI, 2025b), and LLaMA-3.1-8B Instruct FP8 (Grattafiori et al., 2024). For each input (see Table 3), the first prompt requests a neutral version of the text, maintaining its meaning and structure. The second prompt re-injects the target emotion, asking for minimal yet effective stylistic modifications. All outputs are stored for further automatic and human evaluation.

To determine how the processing steps affect fluency, we calculated model perplexities on the original, neutralized, and emotionalized texts via a language-model-based metric (using the GPT-2 (Radford et al., 2019) language model). As shown in Table 4, the original UniC transcripts have the highest perplexity because they are derived from automatic speech recognition (ASR) outputs, which include disfluencies, artifacts, and non-standard punctuation. Neutralization and emotion re-injection produce more structured, readable sentences, resulting in lower perplexity scores. We also measured text length (`len`), for which we observe that LLaMA-3.3-70B tends to produce slightly longer neutralized texts than the original transcripts. However, for most other models, text length decreases in the neutralization step. In contrast, the emotion re-injection step generally increases length across all models, reflecting the use of additional descriptive phrases to convey affect.

5. Automatic evaluation metrics

To systematically assess generative models in their ability to produce emotionally expressive content, we employ metrics that can evaluate such capabilities automatically and with near-human reliability. Although human evaluation offers a nuanced and contextually grounded benchmark for emotional alignment, it is time-consuming, costly, and difficult to scale due to the need for expert annotators and quality control. Therefore, we focus on automatic metrics that enable objective, reproducible, and scalable assessments of emotional generation performance. In this section, we describe the formulas used for the automatic evaluation metrics. We evaluate four types of metrics, detailing their definitions, computational procedures, and roles in assessing different aspects of emotional expression.

5.1. Emotion Probability Score (EPS)

EPS (Dong et al., 2025) evaluates the emotional expressiveness of generated sentences using the bart-large-mnli (Facebook-AI, 2020) model, an open-world classifier capable of zero-shot text classification. They defined three custom labels, *emotionless*, *neutral*, and *emotional*, and we treat each as an independent metric in our experiments to capture the full emotional spectrum of model outputs. Although the distinction between *emotionless* and *neutral* may appear subtle, it follows the prior study's (Dong et al., 2025) formulation to enhance the robustness of emotional assessment. For each label $c \in \{\text{emotionless}, \text{neutral}, \text{emotional}\}$, the corresponding probability score is computed as:

$$\text{EPS}_c = \frac{N_c}{N_{\text{total}}} \tag{1}$$

where N_c is the number of sentences classified as category c, and N_{total} is the total number of evaluated sentences. A higher value of $\text{EPS}_{\text{emotional}}$ indicates stronger emotional expressiveness, while lower values of $\text{EPS}_{\text{emotionless}}$ and $\text{EPS}_{\text{neutral}}$ reflect better performance, as they correspond to reduced neutral or emotionless tendencies in the generated text.

UniC Emotion	Prolific Emotions	Original Text	Neutral Text	Emotional Text
contentment	*joy*: 1 *contentment*: 2 *surprise*: 1 *neutral*: 1	about cannibalism trust me this is a good book we follow our female main character who is a literal psychopath and we follow her as she goes after her ex-lovers kills them and then eats them	about cannibalism this is a book we follow our female main character who is a literal psychopath and we follow her as she goes after her ex-lovers kills them and then eats them	about cannibalism this is a fascinating book we follow our female main character who is a literal psychopath and we follow her with a sense of eager curiosity as she confidently goes after her ex-lovers, kills them with chilling precision, and then eats them. Each page unfolds her unnerving world in a strangely satisfying way.
contentment	*neutral*: 5	noticed that I personally have a preference for like the more like silly light-hearted stories as opposed to beach read which falls more into like the sweeping emotional side of things	noticed that I personally have a preference for the more silly, light-hearted stories as opposed to beach read which falls more into the sweeping emotional side of things	I have a preference for stories that are more light-hearted, the kind that bring a gentle smile or a comforting warmth, as opposed to 'Beach Read,' which falls more into the sweeping emotional side of things.

Table 3: An example annotation from the dataset illustrating the original, neutralized, and emotionally enriched texts generated by GPT-4.1, along with their corresponding emotion labels. The yellow highlights indicate emotionally salient cues and are shown solely for clarification purposes in these two samples.

Model	Neutral		Emotional	
	PPL	len	PPL	len
GPT-4.1	106.02	30.77	72.66	35.11
LLaMA-3.3-70B	84.67	39.37	57.37	38.63
LLaMA-3.1-8B	73.92	30.14	59.26	44.30
Mistral-Nemo-12B	74.83	30.03	55.51	33.27
GPT-4o-mini	63.21	27.22	61.96	35.18

Table 4: Average perplexity (PPL) and number of characters (len) for each model and text type. The original text has a fixed PPL of 108.68 and a length of 33.92 across models.

5.2. Emotion Absolute Score (EAS)

EAS (Dong et al., 2025) quantifies the absolute emotion intensity of each generated sentence using GPT-5-nano with prompt engineering to obtain continuous emotion strength scores (0–100) for six basic emotions. The overall emotion intensity of a sentence is calculated as:

$$\text{EAS} = \sum_{em \in \mathcal{E}} \left(\frac{\text{score}_{em}}{100} \right)^2 \tag{2}$$

where $\mathcal{E} = \{\text{anger}, \text{disgust}, \text{fear}, \text{joy}, \text{sadness}, \text{surprise}\}$ and score_{em} denotes the intensity score assigned to each emotion.

5.3. Emotion Recognition Score (ERS)

ERS is a widely used metric for evaluating generative models in emotion-conditioned text generation. Given a target emotion prompt (e.g., generating a sentence with *anger*), ERS measures the proportion of generated sentences whose predicted emotion label, obtained from an external classifier, matches the intended emotion. Formally, ERS is defined as:

$$\text{ERS} = \frac{N_{\text{matched}}}{N_{\text{total}}} \tag{3}$$

where N_{matched} denotes the number of generated sentences whose predicted emotion matches the target emotion, and N_{total} is the total number of evaluated samples. In our experiments, we employed three different classifiers to compute ERS: the michellejieli (Hartmann, 2022b) as a DistilRoBERTa-base (Sanh et al., 2019) model, and ChatGPT-5-Nano (OpenAI, 2025a) and DeepSeek v3.2 (Liu et al., 2025a) as generative models.

5.4. Emotion Analogy Score (EAnS)

EAnS (Jafari et al., 2025) evaluates how effectively a model reinjects a target emotion into a neutralized sentence while preserving the semantic meaning of the original utterance. This metric relies on analogy reasoning in the embedding space using the BGE-M3 sentence embedder model (Chen et al., 2024a), where emotional relationships are modeled through vector arithmetic between emotion and sentence embeddings. The analogy vector A is defined as:

$$A = E_t - S_t + S_n \tag{4}$$

where E_t and S_t denote the embeddings of the target emotion and its corresponding emotional sentence, E_n and S_n denote the embeddings of the neutral emotion and its corresponding neutralized sentence, and S_o refers to the original sentence before neutralization. To evaluate emotional alignment, we compute three sets of relational metrics:

$$\begin{aligned} R1_c &= \cos(E_n, E_t), & R1_m &= \|E_n - E_t\|_1 \\ R2_c &= \cos(E_n, A), & R2_m &= \|E_n - A\|_1 \\ R3_c &= \cos(S_o, S_t), & R3_m &= \|S_o - S_t\|_1 \end{aligned} \tag{5}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity and $\|\cdot\|_1$ represents the Manhattan distance. In the evaluation, higher cosine similarity values ($R1_c$, $R2_c$, $R3_c$) indicate stronger emotional and semantic alignment, whereas lower Manhattan distances ($R1_m$, $R2_m$, $R3_m$) reflect closer embedding proximity.

In total, our automatic evaluation framework comprises 13 distinct metrics: three derived from EPS

(emotionless, neutral, emotional), one from EAS (overall emotional intensity), three from ERS (corresponding to three classifiers), and six from EAnS (cosine similarity and Manhattan distance across different relational settings). These metrics jointly capture various dimensions of emotional expressiveness and semantic preservation.

6. Human annotation of the generated texts

To evaluate the emotional aptitude and perceived quality of texts generated by the five LLMs, we conducted a systematic human annotation process for both neutralization and emotion re-injection steps. This provides a crucial benchmark for assessing emotional expressiveness beyond automated metrics. Due to the time-consuming and costly nature of human evaluation, we applied it to a representative subset of the data. We then computed the correlation between human evaluation results and various automatic metrics to identify which metric best aligns with human perception. Human evaluation measures whether the generated texts convey the intended emotion, serving as a human-annotated counterpart to the **ERS** metric (see Subsection 5.3). Unlike **ERS**, predictions are made by human annotators, who identify the perceived emotion of each sentence and compare it to the target emotion specified in the generation prompt.

6.1. Subset for manual labeling

As detailed in Section 4, we instructed five distinct LLMs to generate emotionally conditioned transcripts for each of the 964 samples in the UniC dataset, resulting in a total of $5 \times 964 = 4{,}820$ generated texts. Conducting a full-scale human evaluation on this entire set would be prohibitively time-consuming and costly. Therefore, we adopted a two-stage selection p

rocess to obtain a representative and reliable subset for manual evaluation. First, since the UniC dataset comprises 18 videos, we selected one sample from the beginning and one from the end of each video for every emotional label, ensuring diversity across both content and emotion categories. Second, we prioritized samples with high-quality ground-truth annotations. As mentioned before, a small sample of the UniC dataset, i.e., two videos with a total of 61 samples, was labeled by three annotators. We selected those samples with an agreement score (Equation 6) greater than 0.6 on the text modality annotation, leading to 55 samples meeting this reliability threshold. Applying these two criteria yielded 277 high-quality samples out of the original 964. Consequently, across all five LLMs, this resulted in $277 \times 5 = 1{,}385$ generated samples for human evaluation. As shown in Table 5, the correlations (using Pearson) between the automatic evaluation metrics computed on the 277-sample subset and the full dataset are consistently high across all five LLMs. This strong alignment indicates that our subset selection method effectively preserves the characteristics of the entire dataset.

Metric Group	Metric Name	Correlation
EPS	emotionless	0.9766
	neutral	0.9826
	emotional	0.9956
EAS	GPT-5-nano	0.9102
ERS	GPT-5-nano	0.9950
	DeepSeek v3.2	0.9450
	Michellejieli	0.8692
EAnS	R1_m	0.9976
	R2_m	0.9150
	R3_m	0.9997
	R1_c	0.9985
	R2_c	0.8959
	R3_c	0.9996

Table 5: Correlation between automatic evaluation metrics computed on the selected 277-sample subset and the full dataset across five LLMs, demonstrating the representativeness of the subset.

The neutralization process is not the main component of the pipeline, so we annotated a representative subset of the 277 samples. Specifically, we selected 12 samples per emotion category from the 277 samples. Given 7 emotion categories, this resulted in: $12 \times 7 = 84$ samples. Across the 5 models, this yields a total of: $84 \times 5 = 420$ samples for neutralization annotation.

6.2. Annotation details

We employed the Prolific[1] crowdsourcing platform to collect annotations. Human evaluation of neutralization was conducted on an 84-sample subset per LLM to assess the effectiveness of emotion neutralization, to which we added a set of attention-check questions designed to ensure annotator reliability. These attention checks were straightforward, for example: “Please choose the emotion *confusion*.” Any annotator who failed these checks was excluded from the final dataset. Each LLM dataset contained 10 attention-check questions, resulting in 94 items per model. Annotators were informed that each text was an automatically neutralized version of an originally emotional sentence and were asked to identify the perceived emotion. Selecting *neutral* indicates successful neutralization, whereas choosing any other emotion suggests a failure in the neutralization process. Five annotators independently annotated each sample. In total, across

[1] www.prolific.com

Model	Krippendorff's α	Fleiss' κ
GPT-4.1	0.2918	0.2903
LLaMA-3.3-70B	0.3023	0.3008
LLaMA-3.1-8B	0.2851	0.2835
Mistral-Nemo-12B	**0.3924**	**0.3911**
GPT-4o-mini	0.2268	0.2252

Table 6: Inter-annotator agreement (IAA) scores for emotion neutralization evaluation across five different LLMs.

all five LLMs, the annotation process covered 475 samples provided by 25 annotators.

In the human annotation for the generated emotional texts, each LLM contributed 277 samples, to which we again added a set of attention-check questions designed to ensure annotator reliability. These attention checks were similar to those for the neutralization part. Each LLM dataset contained 28 attention-check questions, resulting in 305 items per model. Annotators were asked to select the most appropriate emotion label for each text from the following categories: *Confusion*, *Contentment*, *Disappointment*, *Disgust*, *Joy*, *Neutral*, and *Surprise*. Each annotator labeled 50 samples, and every sample received five independent annotations. Thus, for each model, we recruited 35 unique annotators to complete the 305-item set. In total, across all five LLMs, the annotation process covered 1,525 samples provided by 175 annotators. All annotators were native English speakers residing in either the United States or the United Kingdom.

6.3. Results on the annotated dataset

The details of each human evaluation are described in the following sections.

6.3.1. Annotation results for neutralized texts

To assess the reliability and consistency of the annotations, we computed two robust inter-annotator agreement (IAA) metrics: Krippendorff's α (Krippendorff, 2018) and Fleiss' κ (Fleiss, 1971). Unlike simple percentage agreement, both metrics correct for chance agreement and therefore provide a more conservative estimate of annotation reliability. The IAA results for all evaluated models are shown in Table 6. Overall, the agreement scores indicate *fair agreement* across models. Higher values of α and κ reflect greater consistency among annotators in identifying neutralized text, whereas lower values suggest residual emotional signals or ambiguity in the generated outputs.

Table 7 summarizes the performance of different models on the emotion neutralization task over 84 annotated samples. In the ideal (best-case) scenario, all outputs should be classified as *neutral*, since the models were explicitly instructed to neutralize the emotional content of the input texts. However, after conducting human evaluation, we found that a subset of the generated outputs still conveyed residual emotions and were therefore annotated as non-neutral. For F1-score computation, the ground-truth labels consist of 84 instances of the *neutral* class (i.e., the true values assume all samples should be neutral). The predicted labels correspond to the *majority emotion* assigned by the human annotators to each generated neutralized text.

Among the evaluated models, *LLaMA-3.3-70B* achieves the highest F1-score (0.95), indicating a strong ability to remove emotional cues and produce genuinely neutral outputs, with very few residual emotional mistakes. *LLaMA-3.1-8B* follows with an F1-score of 0.83, demonstrating robust performance despite its smaller model size. *GPT-4.1* and *Mistral-Nemo-12B* achieve comparable F1-scores (0.78), suggesting similar neutralization behavior across proprietary and open-source models. Finally, *GPT-4o-mini* records the lowest F1-score (0.77), reflecting comparatively weaker neutralization performance and a higher number of non-neutral outputs. Overall, these results demonstrate that instruction-tuned open-source models, particularly larger variants, can achieve strong and competitive performance in the neutralization stage of the proposed pipeline.

6.3.2. Annotation results for generated emotional texts

To evaluate the quality and consistency of the annotations, we first measured a simple majority agreement score, defined as:

$$A = \frac{f_{\text{max}}}{N} \tag{6}$$

where A represents the agreement ratio, f_{max} is the frequency of the majority-selected emotion label, and $N = 5$ corresponds to the total number of annotators for each sample. This measure reflects the proportion of annotators who agreed on the most frequently chosen label, providing a direct estimate of labeling consensus. Table 8 summarizes the agreement scores across different emotion categories and models.

Overall, the agreement levels vary notably across emotions and models. *Disappointment* consistently shows the highest agreement (ranging from 0.75 to 0.88 across models), indicating that annotators found this emotion relatively easy to identify. Emotions like *neutral* and *confusion* also exhibit relatively high agreement, suggesting lower ambiguity in these categories. By contrast, *contentment* and *surprise* yield lower agreement scores, reflecting their more subjective and context-dependent

Model	F1	Disappointment	Disgust	Confusion	Neutral	Contentment	Joy	Surprise
GPT-4.1	0.7826	3	5	9	54	3	3	7
LLaMA-3.3-70B	**0.9500**	2	2	1	76	2	0	1
LLaMA-3.1-8B	0.8333	3	2	3	60	3	7	6
Mistral-Nemo-12B	0.7826	8	4	5	54	3	3	7
GPT-4o-mini	0.7737	11	0	6	53	6	3	5

Table 7: Emotion neutralization performance of different models measured using F1-score. The per-emotion columns report the number of samples annotated with each dominant emotion by human evaluators. Since all inputs were expected to be neutral after processing, non-neutral categories represent emotional mistakes made by the models.

nature. We also evaluated annotation agreement on the generation output derived from the high-agreement 55-sample subset. Most models indeed show improved or stabilized agreement scores. For instance, GPT-4.1 improves from 0.7235 to 0.7600 in the overall agreement score after filtering, which confirms that this subset better represents clear-cut emotional expressions. Similar improvements are observed for LLaMA-8B, GPT-4o-mini, and Mistral-12B.

To further validate the consistency of the annotations, we computed two robust inter-annotator agreement metrics: Krippendorff's α and Fleiss' κ. As shown in Table 9, α and κ values fall between 0.38 and 0.45, which indicates a moderate level of agreement among annotators. These values are typical in emotion annotation tasks (Du et al., 2025), which inherently involve subjective interpretation. Notably, GPT-4.1 yields the highest agreement scores ($\alpha = 0.4497$, $\kappa = 0.4469$), suggesting that the emotional cues in its outputs were slightly more consistently perceived by annotators compared to the other models.

Finally, we evaluated the emotion classification performance of the models based on human-labeled annotations, using F1-score as the primary evaluation metric. As shown in Table 10, **GPT-4.1** consistently achieves the best performance across most emotion categories, with a total weighted F1-score of 0.74. This aligns with its higher agreement metrics, suggesting that the emotional expressions it generates are clearer and more consistently interpretable by human annotators. Interestingly, **LLaMA-8B** demonstrates competitive performance in specific categories such as *disappointment* and *disgust*, achieving F1-scores as high as 0.88. However, its overall performance (0.63) remains lower than GPT-4.1 due to weaker results in other emotions, including *contentment* and *joy*. The lowest overall performance is observed for **GPT-4o-mini** (0.58), indicating that smaller models may generate more ambiguous or less consistent emotional content, making the annotation task more challenging. Moreover, when models are explicitly instructed to generate text with a specific target emotion (e.g., *joy*), and human annotators confirm that the generated emotion aligns with the target, the resulting F1-score tends to increase. This demonstrates the positive impact of prompt specificity on both emotional clarity and annotator agreement. In other words, more controlled generation conditions lead to more consistent emotional interpretation, which is reflected in higher human evaluation scores.

In summary, the results indicate that *disappointment* and *confusion* are the most consistently annotated emotions, showing both high agreement and high model performance. Filtering low-agreement samples leads to improved consistency across all models, which suggests that the subset represents clearer emotional signals. GPT-4.1 demonstrates the strongest alignment between annotators and model output, followed by LLaMA-70B, while smaller models such as GPT-4o-mini tend to produce more ambiguous emotional content, reflected in lower agreement and lower F1 scores.

7. Experiments and results

Our evaluation process consists of two main stages. In the first stage, we determine the most reliable automatic evaluation metric by measuring its correlation with human judgments on the manually annotated subset of 277 samples. The metric that demonstrates the strongest agreement with human assessment is then used in the second stage to automatically evaluate the entire dataset and to identify the best-performing model overall.

7.1. Selecting the automatic evaluation metric

Table 11 presents the results of human evaluation (using the ERS metric) and several automatic metrics across five models, along with their correlations with human evaluation. A higher positive correlation indicates closer alignment with human judgment, while negative correlations suggest the opposite. A key finding is that using a strong language model as an evaluator does not guarantee agreement with human assessment. Both EAS and ERS rely on GPT-5-nano, yet their correlations differ sharply: EAS shows a negative correlation, whereas ERS exceeds 0.95, demonstrating that evaluation design and metric formulation matter

Emotion	GPT-4.1		LLaMA-3.3-70B		LLaMA-3.1-8B		Mistral-Nemo-12B		GPT-4o-mini	
	277	55	277	55	277	55	277	55	277	55
Neutral	0.7260	**0.7913**	**0.7412**	0.6500	0.6327	0.7333	0.6348	0.5846	0.6870	0.6545
Disappointment	0.8156	**0.8833**	0.7831	0.8600	**0.8429**	0.8571	0.7803	0.8769	0.7529	0.8154
Contentment	0.6053	0.6000	0.5697	0.5400	0.6000	**0.6400**	0.6059	0.5667	**0.6077**	**0.6400**
Disgust	0.7100	0.7000	0.7273	0.6667	0.7308	0.7500	**0.8000**	**0.8000**	0.6444	0.4000
Confusion	**0.7875**	**0.8000**	0.7647	**0.8000**	0.7111	0.7000	0.6632	**0.8000**	0.7375	**0.8000**
Joy	0.7091	0.6286	0.6737	0.5750	0.7486	**0.8250**	**0.7507**	0.7529	0.6701	0.7294
Surprise	0.6455	**0.8000**	0.6381	0.6000	0.5793	0.5600	**0.6632**	0.6667	0.6308	0.5714
Total	**0.7235**	**0.7600**	0.7083	0.6582	0.7098	0.7454	0.7126	0.7200	0.6859	0.7018

Table 8: Comparison of annotation agreement across models and emotion categories. The left subcolumns show results for all 277 samples, while the right subcolumns report results for the 55-sample subset selected using the high-agreement threshold described in Subsection 6.1. Boldface numbers indicate the highest agreement value for each emotion within a given dataset, highlighting which model performs best for that specific emotion.

Model	Krippendorff's α	Fleiss' κ
GPT-4.1	**0.4497**	**0.4469**
LLaMA-3.3-70B	0.4261	0.4232
LLaMA-3.1-8B	0.4394	0.4364
Mistral-Nemo-12B	0.4451	0.4422
GPT-4o-mini	0.3841	0.3807

Table 9: Inter-annotator agreement scores for emotion re-injection across models using Krippendorff's α and Fleiss' κ.

Emotion	GPT-4.1	LLaMA-70B	LLaMA-8B	Mistral-12B	GPT-4o-mini
Neutral	**0.81**	0.79	0.63	0.63	0.52
Disappointment	0.85	0.82	**0.88**	0.82	0.84
Contentment	**0.60**	0.55	0.39	0.47	0.40
Disgust	0.71	0.77	**0.88**	0.63	0.70
Confusion	0.85	**0.94**	0.86	0.72	0.85
Joy	**0.66**	0.56	0.41	0.50	0.40
Surprise	**0.59**	0.55	0.49	0.45	0.47
Total (Weighted F1)	**0.74**	0.71	0.63	0.61	0.58

Table 10: Human evaluation performance for emotion re-injection across models for the 277-sample dataset.

more than model strength. Among all evaluated metrics, ERS computed with the DeepSeek v3.2 emotion recognizer achieves the highest correlation with human evaluation (0.976), followed by ERS with GPT-5-nano (0.954). It is important to note that GPT-5-nano is a lightweight variant of the GPT-5 model, designed to be smaller and more computationally efficient. As a result, it may have reduced capacity compared to larger models, which reasonably explains why its reliability is slightly lower than that of DeepSeek v3.2 in this evaluation. Metrics such as R3_m show moderate positive correlation (0.655), while EPS and EAS variants yield negative correlations. Therefore, ERS with DeepSeek v3.2 is selected as the most reliable automatic metric for large-scale evaluation.

7.2. Results on the entire dataset

Following the metric selection described in the previous section, two evaluators show the highest correlation with human judgment: ERS computed with DeepSeek v3.2 (open-source) and ERS computed with GPT-5-nano (closed-source). We therefore apply both evaluators to assess model performance on the entire dataset. Table 12 reports the ERS scores of five LLMs under these two evaluation settings. Under GPT-5-nano evaluation, GPT-4.1 achieves the highest ERS score (0.7396). In contrast, when evaluated with DeepSeek v3.2, LLaMA-3.3-70B obtains the best performance (0.6556), slightly outperforming GPT-4.1. Tables 13 and 14 provide a detailed breakdown of emotion generation performance for the best-performing model under each evaluator. Table 13 presents the results for GPT-4.1 under GPT-5-nano evaluation, while Table 14 reports the results for LLaMA-3.3-70B under DeepSeek v3.2 evaluation. Overall, GPT-4.1 shows higher accuracy (0.740 vs. 0.656) and a stronger weighted F1 score (0.746 vs. 0.658). Across most emotion categories, GPT-4.1 also demonstrates a more balanced precision-recall trade-off, particularly for high-support classes such as Contentment, Disappointment, and Neutral.

8. Conclusion

In this work, we presented a robust and extensible benchmark designed to evaluate automatic evaluation metrics for emotion generation, as well as to assess the performance of LLMs on this task. The framework currently includes 13 automatic evaluation metrics and five LLMs, but its design allows for seamless integration of additional metrics and models without the need for further human annotation. A key finding of our study is that using a strong language model as an evaluator does not guarantee reliable alignment with human judgment.

Model	Human Eval (F1)	EPS			EAS	ERS			EAnS					
		emotionless	neutral	emotional	GPT-5-nano	GPT-5-nano	DeepSeek v3.2	Michellejieli	R1_m	R2_m	R3_m	R1_c	R2_c	R3_c
GPT-4.1	**0.736**	0.155	0.281	0.564	0.586	**0.773**	0.671	0.596	**11.08**	6.445	**10.84**	**0.894**	0.962	**0.900**
LLaMA-3.3-70B	0.711	0.157	0.274	0.569	0.755	0.729	**0.675**	0.639	13.38	**6.42**	13.63	0.846	**0.964**	0.846
LLaMA-3.1-8B	0.627	0.119	**0.199**	**0.681**	**0.818**	0.574	0.563	**0.646**	15.13	6.863	15.32	0.813	0.960	0.809
Mistral-Nemo-12B	0.614	0.138	0.205	0.658	0.771	0.574	0.566	0.628	14.41	7.133	14.73	0.830	0.957	0.824
GPT-4o-mini	0.580	**0.118**	0.238	0.644	0.732	0.585	0.545	0.578	11.75	6.590	13.39	0.884	0.961	0.854
Correlation	1.000	-0.901	-0.784	-0.872	-0.622	0.954	**0.976**	0.150	0.344	0.622	0.655	0.282	0.630	0.618

Table 11: Correlation between human evaluation (F1) and automatic metrics on the 277-sample annotated subset. Higher correlation values indicate stronger agreement with human judgments.

Model	GPT-5-nano	DeepSeek v3.2
GPT-4.1	**0.7396**	0.6027
LLaMA-3.3-70B	0.7199	**0.6556**
LLaMA-3.1-8B	0.5373	0.4959
Mistral-Nemo-12B	0.5384	0.5290
GPT-4o-mini	0.5405	0.5135

Table 12: ERS comparison between DeepSeek-v3.2 and GPT-5-nano evaluators on the entire dataset.

Emotion	Support	Precision	Recall	F1
Confusion	25	0.415	0.880	0.564
Contentment	237	0.920	0.582	0.713
Disappointment	200	0.733	0.905	0.810
Disgust	72	0.776	0.528	0.628
Joy	90	0.485	0.900	0.630
Neutral	325	0.903	0.742	0.814
Surprise	15	0.387	0.800	0.522
Accuracy	964		0.740	
Macro Avg	964	0.660	0.762	0.669
Weighted Avg	964	0.803	0.740	0.746

Table 13: ERS evaluation per emotion using GPT-5-nano as the evaluator. Results correspond to GPT-4.1, the best-performing model under this evaluation setting.

Emotion	Support	Precision	Recall	F1
Confusion	25	0.353	0.720	0.474
Contentment	237	0.882	0.409	0.559
Disappointment	200	0.726	0.860	0.787
Disgust	72	0.624	0.736	0.675
Joy	90	0.397	0.878	0.547
Neutral	325	0.800	0.628	0.703
Surprise	15	0.333	0.600	0.429
Accuracy	964		0.656	
Macro Avg	964	0.588	0.690	0.596
Weighted Avg	964	0.735	0.656	0.658

Table 14: ERS evaluation per emotion using DeepSeek-v3.2 as the evaluator. Results correspond to LLaMA-3.3-70B, the best-performing model under this evaluation setting.

Specifically, both EAS and ERS rely on GPT-5-nano. Yet, their correlation with human evaluation scores differs dramatically: EAS exhibits a negative correlation, while ERS achieves a correlation above 0.94 with human ratings. This sharp contrast highlights that the evaluation design and metric formulation are crucial factors, and not merely the strength of the underlying language model. For scenarios where the use of LLMs as evaluators is not feasible, we also identify practical, lightweight alternatives within the EAnS metric family. The R2_m metric can be used effectively when no original dataset is available, and R3_m provides even better performance when the original dataset is accessible. These options offer accessible solutions for low-resource settings.

9. Future Work

There are several directions for extending this work. First, we plan to increase both the number of language models and the size of the evaluation sample to build a more statistically robust and reliable framework for evaluating automatic metrics. Second, the benchmark can be expanded to support additional modalities such as video and speech, enabling a more comprehensive assessment of multimodal emotion generation systems. Third, future work can explore improving the emotion generation capabilities of language models themselves, aiming to achieve higher and more consistent evaluation scores across different metrics. Finally, we intend to further investigate and refine automatic evaluation metrics, especially those based on LLMs, to improve their alignment with human perception and reduce their variability.

10. Limitations

Despite its strengths, our framework has several limitations. The current evaluation setup focuses solely on textual emotion generation and does not yet support multimodal or interactive scenarios. Also, although our dataset is carefully curated, it may not capture the full range of emotional expressions encountered in real-world applications.

11. Ethical Considerations

This work evaluates large language models that process emotionally sensitive language, which may influence domains such as mental health communication, education, and political discourse. The proposed benchmark is intended solely for research and evaluation and is not designed for clinical, therapeutic, or political decision-making. The evaluation data partially relies on the UniC dataset, which contains transcripts of publicly available YouTube monologues and is distributed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license. Access to the dataset can be requested via its repository[2] and by contacting Quanqi Du[3]. The dataset annotations were collected from 175 crowd workers using the Prolific platform[4], where annotators participate under Prolific's terms of service[5]. Annotators' identities remained anonymized to the researchers, and only randomly assigned Prolific participant IDs were visible. Participation required agreement to Prolific's platform terms by both researchers and annotators, which includes informed consent for participation in research tasks and permission for the resulting annotations to be used for academic research purposes, thereby covering consent and intellectual property usage within the platform's framework. Annotators were compensated according to Prolific's fair payment guidelines; in total, £630 was paid, corresponding to an average hourly rate of approximately £8. Communication channels provided by Prolific allowed annotators to report issues during the annotation process. When annotation quality issues were detected, submissions could be returned for revision with feedback. In cases where technical problems occurred on the researchers' side (e.g., platform or task-related issues), annotators were instructed to report the issue and were compensated when the problem was verified.

Acknowledgments

This research received funding from the Flemish Government under the Flanders Artificial Intelligence Research program (FAIR) (174P07826).

12. Bibliographical References

[2] www.huggingface.co/datasets/LT3/UniC

[3] quanqi.du@ugent.be

[4] www.prolific.com

[5] researcher-help.prolific.com/en/articles/445120-prolific-supplier-vendor-information

Mistral AI. 2024. Mistral nemo. https://mistral.ai/news/mistral-nemo/. Accessed: September 23, 2024.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024b. EmotionQueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176, Bangkok, Thailand. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Celso M De Melo, Peter Carnevale, and Jonathan Gratch. 2010. The influence of emotions in embodied agents on human decision-making. In *International conference on intelligent virtual agents*, pages 357–370. Springer.

Yurui Dong, Luozhijie Jin, Yao Yang, Bingjie Lu, Jiaxi Yang, and Zhi Liu. 2025. Controllable emotion generation with emotion vectors. *arXiv preprint arXiv:2502.04075*.

Du, Quanqi and Labat, Sofie and Demeester, Thomas and Hoste, Veronique. 2025. *UniC: A dataset for emotion analysis of videos with multimodal and unimodal labels*. Springer.

Facebook-AI. 2020. facebook/bart-large-mnli. https://huggingface.co/facebook/bart-large-mnli. Accessed: 2025-06-12.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024.

The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jochen Hartmann. 2022a. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Jochen Hartmann. 2022b. Fine-tuned DistilRoBERTa-base for Emotion Classification. https://huggingface.co/michellejieli/emotion_text_classifier/. Accessed: 2025-05-07.

Sadegh Jafari, Els Lefever, and Véronique Hoste. 2025. Embedding analogies for evaluating emotion in llm-generated utterances. In *28th European Conference on Artificial Intelligence (ECAI 2025)-BEHAIV workshop*.

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

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Weichu Liu, Jing Xiong, Yuxuan Hu, Zixuan Li, Minghuan Tan, Ningning Mao, Chenyang Zhao, Zhongwei Wan, Chaofan Tao, Wendong Xu, et al. 2025b. Longemotion: Measuring emotional intelligence of large language models in long-context interaction. *arXiv preprint arXiv:2509.07403*.

OpenAI. 2025a. Gpt-5 nano. OpenAI model documentation. https://platform.openai.com/docs/models/gpt-5-nano.

OpenAI. 2025b. Introducing gpt-4.1. https://platform.openai.com/docs/models/gpt-4.1. Accessed: 2025-06-04.

OpenAI. 2025c. Introducing gpt-4o-mini. https://platform.openai.com/docs/models/gpt-4o-mini. Accessed: 2025-06-04.

Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ishika Singh, Ahsan Barkati, Tushar Goswamy, and Ashutosh Modi. 2020. Adapting a language model for controlled affective text generation. *arXiv preprint arXiv:2011.04000*.

Haoqin Sun, Xuechen Wang, Jinghua Zhao, Shiwan Zhao, Jiaming Zhou, Hui Wang, Jiabei He, Aobo Kong, Xi Yang, Yequan Wang, et al. 2025. Emotiontalk: An interactive chinese multimodal emotion dataset with rich annotations. *arXiv preprint arXiv:2505.23018*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rasita Vinay, Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2024. Emotional manipulation through prompt engineering amplifies disinformation generation in ai large language models. *arXiv preprint arXiv:2403.03550*.

Shiquan Wang, Ruiyu Fang, Zhongjiang He, Shuangyong Song, and Yongxiang Li. 2025. Emotional support with llm-based empathetic dialogue generation.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2025. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Neural Networks*, page 107901.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. *arXiv preprint arXiv:2205.10237*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.