From Text to Emotion: Unveiling the Emotion Annotation Capabilities of LLMs

Minxue Niu[1]*, Mimansa Jaiswal*[2]*, Emily Mower Provost*[1]

[1]University of Michigan, USA [2]Independent Researcher

sandymn@umich.edu, mimansa.jaiswal@gmail.com, emilykmp@umich.edu

Abstract

Training emotion recognition models has relied heavily on human annotated data, which present diversity, quality, and cost challenges. In this paper, we explore the potential of Large Language Models (LLMs), specifically GPT-4, in automating or assisting emotion annotation. We compare GPT-4 with supervised models and/or humans in three aspects: agreement with human annotations, alignment with human perception, and impact on model training. We find that common metrics that use aggregated human annotations as ground truth can underestimate GPT-4's performance, and our human evaluation experiment reveals a consistent preference for GPT-4 annotations over humans across multiple datasets and evaluators. Further, we investigate the impact of using GPT-4 as an annotation filtering process to improve model training. Together, our findings highlight the great potential of LLMs in emotion annotation tasks and underscore the need for refined evaluation methodologies.
Index Terms: Emotion Recognition, Large Language Models

1. Introduction

Understanding human emotions from written or spoken language is crucial is a key part of studying how computers can interact with us more like humans do. The field has attracted significant research efforts, ranging from word-level analysis [1, 2] to building sophisticated neural networks [3, 4]. Currently, many models demonstrate impressive capabilities in recognizing various human emotions.

The training of emotion models has relied heavily on datasets with human annotations. However, obtaining emotion annotations is challenging due to the rich, ambiguous and subjective nature of emotions [5–7]. The first challenge is to identify the emotion theory that will motivate a particular labeling schema. Common theories include basic emotion theory [8], assigning one or more predefined emotion classes to each sample (categorical labels), and the emotion circumplex theory [9], rating each sample on continuous scales, such as valence and arousal to reflect the emotion's positivity and intensity (dimensional labels). The process of collecting human annotations involves multiple annotators per sample to accommodate subjective interpretations and possible quality issues, with the final label often determined through aggregation methods like majority voting [10] or averaging [11]. Given the large scale of modern datasets, such annotation processes can be both costly and time-consuming. Moreover, the complexity of the label space and the difficulty of quality control further add to the challenges.

Recently, the progress in LLMs brings new alternatives. With remarkable proficiency in language modeling across a wide range of scenarios, LLMs show emerging common sense reasoning capability [12]: they can answer a wide range of natural language reasoning questions through zero- or few- shot prompting, matching or even outperforming supervised models [13–15]. What's more, LLMs display an understanding of human emotion and can respond differently to emotional content [16, 17]. This has inspired research into leveraging LLMs as emotion models to aid emotion annotation processes.

In this work, we comprehensively assess GPT-4's potential to perform emotion annotations in a zero-shot manner. We first measure its emotion recognition performance and find that it performs comparably to established supervised models as baselines, using human annotations as the ground truth. We then reflect on the differences between GPT-perception and human-perception and evaluate how those differences are perceived by a separate set of human evaluators. Surprisingly, we find that human evaluators consistently prefer the GPT-4 annotations over human annotations. These findings raise important open questions about the suitability of conventional "ground truth" concepts and evaluation practices, especially as models begin to approach human-level performance. Further, we analyze how label formats (categorical vs. dimensional) affect GPT-4's performance, and we explore the feasibility of applying GPT-4 as a quality checker for existing annotations. We demonstrate that GPT-4 can identify potentially low-quality annotations and help with curating a cleaner and more efficient training set.

In summary, our research reveals the great potential of utilizing LLMs for emotion annotation tasks, offers new insights into their capabilities across label formats, and highlights the challenges involved in their evaluation. We also release the GPT-4 annotations on the entire GoEmotions dataset, along with our code and prompts[1].

2. Related Work

Affective capabilities of LLMs. Many evaluation studies have shown that LLMs are equipped with emotional intelligence: they are able to derive appraisals of given situations [18], identify the emotions and emotion causes in dialogues [16], and respond with emotional support [16, 17, 19]. Yet, they are generally found to be inferior to humans: a few works that developed benchmarks for assessing emotional intelligence consistently indicate a notable gap in complex emotion reasoning between state-of-the-art LLMs and human performance [17, 20, 21]. There have been a few works that evaluate GPT's zero- or few-shot capability of emotion recognition from text or speech input [22–25]. However, the diversity of emotion label spaces are rarely discussed. Besides, existing works adopt evaluation criteria that rely on automatic metrics against human annotations as the ground truth. In this work, we show that such metrics can

[1]https://github.com/chailab-umich/GPT-4-Emotion-Annotation

Table 1: *Dataset details. label: C-categorical, D-dimensional. The column "Multi" indicates whether it's a multilabel classification task. "Indiv." indicates whether individual annotations on each sample are released.*

Dataset	Domain	Label (d)	Multi.	Indiv.
ISEAR	self reports	C (7)	No	No
SemEval	tweets	C (11)	Yes	No
GoEmotions	reddits	C (28)	Yes	Yes
Emobank	multi-genre	D (3)	N/A	Yes

be biased and may undervalue GPT's effectiveness.

LLMs as data annotators. Despite their remarkable capabilities in various language understanding tasks [14, 15, 26], the high operational costs and impracticality of deployment on edge devices have focus efforts towards using LLMs to enhance annotation processes for training more compact models. GPT has been recognized for its potential in sample annotation [27] and generation [28, 29]. In a closer look, LLMs especially excel at tasks with limited and well-defined label sets [28].

Prompting methods. It is widely acknowledged that LLMs are sensitive to the format and word choices in the prompts [30], making prompts the key factor in the successful application of LLMs. There are two common ways of prompting [31]: cloze prompts, which involve a fill-in-the-blank approach (e.g.,"I feel [X]. I finally got that promotion!"), and prefix prompts, where the model extends a given prompt (e.g., "'I finally got the promotion!' What is the speaker's emotion?"). Given GPT-4's pretraining on generation tasks, our study employs prefix prompts. There have been a lot of work exploring different techniques of prompting that could bring a significant improvement in the models' responses [31–33]. The efficiency of different prompting techniques is not the focus of this paper. We follow the common effective practices without dedicated prompt engineering (details in Section 4.1).

3. Data

We use four publicly available emotion recognition datasets for our analysis, encompassing a variety of label representations and diverse text domains (Table 1). Considering the substantial volume of these datasets, we first select a subset of 500 samples from each for GPT-4 annotation and subsequent analysis.

International Survey on Emotion Antecedents and Reactions (ISEAR) [34] is an outcome of a psychological study aiming to understand the antecedents and reactions to seven basic emotions (joy, fear, anger, sadness, disgust, shame, guilt). It consists of 7.6k samples from firsthand emotional reports in text form. We randomly select 500 samples for our analysis.

SemEval 2017 Task 4 (SemEval) [10] is part of the International Workshop on Semantic Evaluation. It consists of Twitter text samples, each annotated with one or more of 11 emotion classes. Since this dataset is very unbalanced, we conduct weighted sampling to select the 500 samples by applying log inverse frequency weighting to the labels, in order to include more emotion labels in our analysis. If a sample carries multiple emotions, the weighting is determined by the rarest label.

GoEmotions [35] is a comprehensive dataset with 58k samples derived from Reddit comments, designed for fine-grained emotion detection. It is characterized by its extensive range of 27 distinct emotion categories, including admiration, remorse, gratitude, etc. Each sample can be assigned one or more emotion labels, as well as an extra "neutral" option. We also apply log inverse frequency weighting in our selection of 500 samples, to address the label imbalance.

Emobank [11] consists of 10k English sentences balancing multiple genres (newspapers, blogs, etc.). The samples are annotated with dimensional emotion labels in the Valence-Arousal-Dominance (VAD) space on a 5-point scale. We focus on the valence score in this study, as it is most commonly included in related literature [25, 36]. Notably, EmoBank distinguishes between the emotional perceptions of writers and readers [37]; we use the reader's annotations, to be consistent with the perspective of GPT-4. We weight the samples by their log deviation from neutral score, to encourage the inclusion of stronger emotional content. I.e., $w_i = log|V_i - 3|$.

4. Methods

4.1. GPT-4 Prompting

For each of the three emotion classification datasets, we collect two sets of GPT-4 annotations. In the first set of annotations, we ask GPT-4 to conduct classification annotations by making selections from a pre-determined set of emotion classes. Informed by the common prompting techniques detailed in Section 2, we follow an instruction-based prompting method, which is consistent with the tasks given to human annotators. We try to give GPT-4 similar instructions as those given to humans, based on the descriptions in the GoEmotions paper [35]. Additionally, we set up a persona in the beginning, which has been found to be effective in our preliminary experiments. As an example, the prompt we use for multi-label classification datasets (GoEmotions and SemEval) is shown below.

"You are an emotionally-intelligent and empathetic agent. You will be given a piece of text, and your task is to identify all the emotions expressed by the writer of the text. You are only allowed to make selections from the following emotions, and don't use any other words: [Options]. Only select those ones for which you are reasonably confident that they are expressed in the text. If no emotion is clearly expressed, select 'neutral'. Reply with only the list of emotions, separated by comma."

We make minimal modifications as needed for other tasks/datasets and all prompts we use across datasets/tasks can be found in our released code.

We then ask GPT-4 to freely generate descriptors of the expressed emotion, without a given range of options. We compare the generated descriptors with the classification results to understand how the granularity of emotion labels affect GPT-4's performance (Section 5.1). For Emobank, we use a similar prompt with the expected response being a integer number from 1 to 5, indicating the perceived valence of the expressed emotion. Using these prompts, we obtain GPT-4 emotion annotations on the 2000 samples selected from four datasets. We additionally obtain classification annotations on all of the GoEmotions dataset using GPT-4 for our model training analysis (Section 5.2).

4.2. Automatic Evaluation Metrics

Following common approaches in previous work, we evaluate GPT-4's performance on two aspects: 1) agreement with human annotations [28, 36], and 2) potential to improve model performance when GPT-4's annotations are used as training data [24, 25] to train smaller models, in this case implemented by fine-tuning BERT. For classification, we use Unweighted Average Recall (UAR) and Macro-averaged F-1 (Macro-F1) scores as the metrics. UAR measures a model's ability to correctly identify instances of each class with equal importance,

while Macro-F1 assesses the balance between precision and recall for all classes. For regression, we use Pearson Correlation Coefficient (PCC) to measure the strength and direction of the linear correlation, and Mean Absolute Error (MAE) to reflect the average error magnitude.

4.3. Supervised model: Finetuned BERT

We finetune BERT [38] models on the full training set of each dataset to serve as a supervised baseline. BERT is one of the most commonly used models for text classification tasks [39] and has been used as a benchmark model for the GoEmotions dataset [35]. Besides, its smaller size means it can be run on a single GPU, so we also use it as our base model when comparing the training efficiency of different annotation sources.

We use the same finetuning settings across all experiments: we use the "bert-base-uncased" model implemented in the transformers library, which has 110M parameters. We add a linear layer on top of the base model, and finetune the whole model on a training set with an AdamW optimizer and learning rate = 1e-5. We optimize a Binary Cross Entropy loss for multi-label classification tasks, a Cross Entropy loss for the single-label classification task, and a Mean Squared Error loss for the regression task. We train the model for 10 epochs, and use the model with best performance on a validation set for testing. For the regression task, we find the model not yet converged after 10 epochs, so we train the model for 30 epochs.

4.4. Human Evaluation

Human annotations often contain inaccuracies [40], thus metrics based solely on human annotations can be biased. Therefore, we conduct a human evaluation study on samples where GPT-4 and the human evaluators do not agree, aiming to incorporate human perspectives into our evaluation.

We recruited four students from the University of Michigan as our evaluators, aged between 19 to 28 and including two females. They were presented with annotations from two sources (i.e., human vs. GPT-4 classification or GPT-4 classification vs. generation) without identification, and were asked to choose the one which they thought "better and more accurately describes the emotion expressed in the text". Each sample was evaluated by two evaluators, who were given an option to indicate uncertainty on each sample. For the classification tasks (ISEAR, SemEval and GoEmotions), we first found all samples that were annotated with disjoint sets of labels by the two sources. Note that we did not adjudicate annotations if they contained overlapping emotion labels as the differences can be subtle (e.g., "anger" vs. "anger, annoyance"). The annotations were randomized and mixed from three different datasets to reduce the likelihood that evaluators could recognizing patterns associated with a specific source. For the regression task (Emobank), it is harder for evaluators to decide whether a given number is a more or less accurate valence rating for a given sample, especially when the ratings are close. Thus, we adopted a relative evaluation schema [41]. We found pairs in disagreement where one annotation source assigns sample A a significantly (> 1 standard deviation) higher rating than sample B, while the other indicates reversed significance. We asked evaluators to indicate which of the two samples in each pair should have the higher valence. The order of the samples was randomized.

5. Results

5.1. GPT-4 Zero-shot Performance

We first compare GPT-4 and human classification annotations, with a focus on their disagreements. We visualize the disagreements on the ISEAR dataset as a confusion matrix in Figure 1 (we select ISEAR because it has the fewest number of classes and is the clearest to show). GPT-4 aligns well with human annotations on most samples, as indicated by the numbers on the diagonal. It's worth noting that confusions mostly happen among similar emotions, and the confusion between a positive emotion and a negative one is rare. Further, it shows that the confusion between classes is not symmetric, indicating some systematic differences between human and GPT-4 annotations. For example, GPT-4 tends to perceive more shame than guilt (18), but seldom marks human-perceived shame as guilt (3).

Table 2: *GPT-4 zero-shot vs. BERT finetuned performance across four dataset. Better performances are in bold.*

	Macro-F1 ↑		**UAR** ↑	
	GPT-4	*BERT*	*GPT-4*	*BERT*
ISEAR	**0.739**	0.726	**0.747**	0.727
SemEval	0.511	**0.548**	0.476	**0.495**
GoEmotions	0.375	**0.521**	**0.485**	0.469
	PCC ↑		**MAE** ↓	
	GPT-4	*BERT*	*GPT-4*	*BERT*
Emobank	**0.764**	0.321	0.645	**0.442**

We then quantitatively evaluate the zero-shot efficacy of GPT-4, and compare its performance to a BERT model fine-tuned to predict the human evaluations. Our findings are in line with prior work in this space [28] that the two approaches perform comparably, and GPT-4 performs slightly better than BERT on the easier 7-class classification dataset ISEAR, but was more challenged on the multi-label classification datasets SemEval and GoEmotions (Table 2).

However, the subsequent human evaluation reveals a different trend and suggests that the automatic metrics may have underestimated GPT-4 performances. As shown in Figure 2 (a) with the colored bars representing the ratio of human preference obtained on each annotation source (Human vs. GPT-4 classification), human evaluators prefer labels from GPT-4 on more samples than those from human annotators, consistently across datasets: ISEAR 62.3%, SemEval 68.2%, GoEmotions 71.1%. This trend holds for each individual annotator, ranging from 64.1% to 71.2%.

Further, Figure 2 (b) shows that GPT-4 generated emotion descriptions are preferred to GPT-4 classification annotations by human evaluators, indicating that without the pre-defined classes as a restriction, GPT-4 generates emotion descriptions that were more often preferred by human evaluators. This trend is more significant when the label set is small, like ISEAR (7-classes 65.4%) and SemEval (11 classes, 73.8%), compared to GoEmotions (28 classes, 55.2%). This comparison indicates that it's beneficial to have a larger label space, which is more likely to encompass the precise emotion labels needed for accurate annotation. The results in Figure 2 (a) highlight the proficiency of GPT-4 in navigating a wide range of labels, further demonstrating its utility in complex emotion recognition tasks.

On the valence regression task, GPT-4 significantly outperforms fine-tuned BERT when measured by PCC, but it has a larger MAE (Table 2). The large MAE can be explained by the highly centralized distribution of human annotations (standard deviation for human evaluations was 0.54, vs. 1.16 for GPT-4) and the fact that GPT-4 predicts integer-valued numbers while the human evaluations are continuous (e.g., averages of multiple evaluators). However, the large PCC value (0.764) indicates that GPT-4 can identify relative emotional valence. Human evalua-

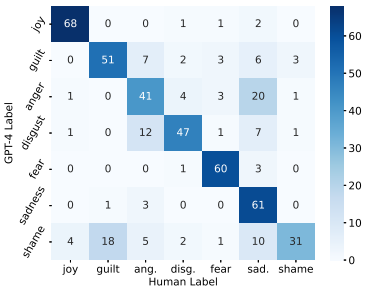
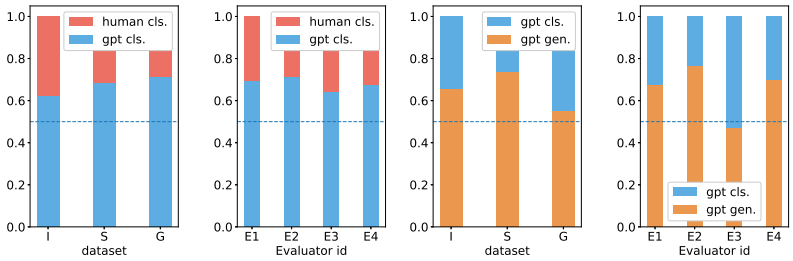


Figure 1: *Disagreements between human and GPT annotations on ISEAR.*



(a) *Human vs. GPT-4 classification.* (b) *GPT-4 classification vs. generation.*

Figure 2: *Human preference ratio comparing human annotations, GPT-4 classification annotations and GPT-4 generation annotations on emotion classification tasks.*

tion also finds an overall 56% agreement with GPT-4 rather than the original human annotations.

5.2. Impact on Model Training

We then investigate whether the labels resulting from GPT-4 classification annotations can be used to train emotion recognition models. We focus on the GoEmotions dataset for this study, using its original train/val/test split.

We compare the performance of a BERT model when it is fine-tuned on the whole training set (N_{train} = 42,278) with human annotations to one trained using the GPT-4 annotations. Additionally, we downsample the training data, retaining only data where the original human evaluations agree with the GPT-4 annotations. We refer to this set as the filtered human set (Human-F, N_{train} = 19,130). Note that this set potentially contains easier samples, compared to both the original human evaluations and GPT-4 annotation labels, because ambiguous samples are more likely to receive different annotations from human and GPT-4, and would thus be filtered out. We test the model on the original human evaluation data (N_{test} = 5,283), the GPT-4 annotations (N_{test} = 5,283), and the Human-F test set (N_{test} = 2,409). We add an extra test set that we refer to as the "adjudicated" test set ($N_{test} = 405$), which is a subset of the 500 samples used in the human preference evaluation experiment. The set is first populated with samples that have overlapping labels from the original human evaluations and GPT-4 ($N = 217$), and either the human or the GPT-4 label is selected by random. The remaining samples exhibit disagreement between the two sources. We select the subset of samples where humans exhibited a clear preference for either the original human evaluation or GPT-4 label[2] as the final label ($N = 188$). The performance on the adjudicated test set is our main metric, because the annotations have been adjudicated and are considered to be more reliable than the raw human or GPT-4 annotations.

In Table 3, the models perform most accurately when trained and tested on the same type of annotation. When models are trained on human annotations and tested with GPT-4 annotations (and vice versa) there are notable performance decreases. This indicates that the models learn a systematic difference between human and GPT-4 annotations, which echos our findings in Section 5.1. On the adjudicated test set, we find that the model trained on GPT-4 annotations outperformed the model trained on human annotations by a large margin (0.524 vs. 0.392, respectively), again pointing to the systematic differences between the two annotation sources. We find that models trained on the filtered subset of the original human evaluations estimate the labels of the adjudicated data more accurately than models trained on the full set of the original human evaluations (0.430 vs. 0.392, respectively). This result is notable given that the Human-F set is only 45% of the size of the original Human set ($N = 19,130$ vs. $N = 42,278$, respectively).

[2]Samples where the human evaluators did not agree on the preferred annotation were not included in this sample, 19% of the samples.

Table 3: *Performance (Macro-F1) of models trained and tested on different combinations of annotations. We show the best performance on each test set (per column) in bold.*

Train \ Test	Human	GPT-4	Human-F	Adjudicated
Human (42k)	**0.486**	0.304	0.568	0.392
GPT-4 (42k)	0.343	**0.517**	0.533	**0.524**
Human-F (19k)	0.478	0.367	**0.617**	0.430

6. Discussion, Limitations and Conclusion

In this work, we evaluate GPT-4's emotion recognition capability and find that its zero-shot performance is comparable to supervised models. Our human evaluation study reveals that GPT-4 annotations are preferred to human annotations by our human evaluators, and GPT-4 is good at handling a wide range of options in emotion classification tasks. We also show that models trained on GPT-4 annotations are subsequently better at predicting the labels amongst the adjudicated subset of data. These results highlight the potential of LLMs to be applied in emotion recognition applications.

Several factors may contribute to the observed preference for GPT-4 annotations. First, humans make mistakes, and the increased cognitive load on more complex label spaces could have increased the vulnerability [42]. Additionally, given the inherent subjectivity and ambiguity of emotion annotations [7], different preferences could indicate variations in annotation perspectives or reflect a lack of diversity in the annotation process. Further exploration is needed to identify the underlying reason. Our findings emphasize the need to reconsider conventional notions of "ground truth" and explore novel evaluation metrics as LLMs approach and surpass human-level performance.

7. References

[1] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceed-*

ings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers), 2018, pp. 174–184.

[2] L. P. Hung and S. Alias, "Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 27, no. 1, pp. 84–95, 2023.

[3] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2937–2987, 2020.

[4] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 49–67, 2021.

[5] W. Wu, C. Zhang, and P. C. Woodland, "Estimating the uncertainty in emotion attributes using deep evidential regression," *arXiv preprint arXiv:2306.06760*, 2023.

[6] H. Tran, I. Falih, X. Goblet, and E. M. Nguifo, "Do multimodal emotion recognition models tackle ambiguity?" in *Proceedings of the 2nd Workshop on People in Vision, Language, and the Mind*, 2022, pp. 6–11.

[7] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[8] P. Ekman *et al.*, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[9] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[10] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," *arXiv:1912.00741*, 2019.

[11] S. Buechel and U. Hahn, "EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," in *Proceedings of the 15th Conference of EACL*, Valencia, Spain, Apr. 2017, pp. 578–585.

[12] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv:2212.10403*, 2022.

[13] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. Huang, "A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 431–469.

[14] T. Brown *et al.*, "Language models are Few-Shot learners," May 2020.

[15] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are Zero-Shot learners," Sep. 2021.

[16] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin, "Is ChatGPT equipped with emotional dialogue capabilities?" Apr. 2023.

[17] J.-T. Huang, M. H. Lam, E. J. Li, S. Ren, W. Wang, W. Jiao, Z. Tu, and M. R. Lyu, "Emotionally numb or empathetic? evaluating how LLMs feel using EmotionBench," Aug. 2023.

[18] A. N. Tak and J. Gratch, "Is GPT a computational model of emotion?" in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[19] C. Li, J. Wang, K. Zhu, Y. Zhang, W. Hou, J. Lian, and X. Xie, "Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus," *arXiv e-prints*, pp. arXiv–2307, 2023.

[20] X. Wang, X. Li, Z. Yin, Y. Wu, and J. Liu, "Emotional intelligence of large language models," *Journal of Pacific Rim Psychology*, vol. 17, p. 18344909231213958, Jan. 2023.

[21] S. Sabour, S. Liu, Z. Zhang, J. M. Liu, J. Zhou, A. S. Sunaryo, J. Li, T. M. C. Lee, R. Mihalcea, and M. Huang, "EmoBench: Evaluating the emotional intelligence of large language models," 2024.

[22] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "Bias in emotion recognition with ChatGPT," 2023.

[23] S. Feng, G. Sun, N. Lubis, C. Zhang, and M. Gašić, "Affect recognition in conversations using large language models," Sep. 2023.

[24] S. Latif, M. Usama, M. I. Malik, and B. W. Schuller, "Can large language models aid in annotating speech emotional data? uncovering new frontiers," Jul. 2023.

[25] Z. Zhang, L. Peng, T. Pang, J. Han, H. Zhao, and B. W. Schuller, "Refashioning emotion recognition modelling: The advent of generalised large models," Aug. 2023.

[26] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang, "Text classification via large language models," *arXiv e-prints*, p. arXiv:2305.08377, May 2023.

[27] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 120, no. 30, p. e2305016120, Jul. 2023.

[28] B. Ding, C. Qin, L. Liu, Y. K. Chia, B. Li, S. Joty, and L. Bing, "Is GPT-3 a good data annotator?" in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jul. 2023, pp. 11 173–11 195.

[29] S. Thapa, U. Naseem, and M. Nasim, "From humans to machines: Can ChatGPT-like LLMs effectively replace human annotators in NLP tasks?" https://workshop-proceedings.icwsm.org/pdf/2023_15.pdf, 2023.

[30] M. Loya, D. A. Sinha, and R. Futrell, "Exploring the sensitivity of llms' decision-making capabilities: Insights from prompt variation and hyperparameters," *arXiv:2312.17476*, 2023.

[31] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Jan. 2023.

[32] M. Binz and E. Schulz, "Using cognitive psychology to understand GPT-3," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 120, no. 6, p. e2218523120, Feb. 2023.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[34] H. G. Wallbott and K. R. Scherer, "How universal and specific is emotional experience? evidence from 27 countries on five continents," *Social Science Information*, vol. 25, no. 4, 1986.

[35] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4040–4054.

[36] T. Feng and S. Narayanan, "Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting," Sep. 2023.

[37] S. Buechel and U. Hahn, "Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation," in *Proceedings of the 11th linguistic annotation workshop*, 2017, pp. 1–12.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[39] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing bert against traditional machine learning text classification," *arXiv preprint arXiv:2005.13012*, 2020.

[40] H. O. Ikediego, M. Ilkan, A. M. Abubakar, and F. V. Bekun, "Crowd-sourcing (who, why and what)," *International Journal of Crowd Science*, vol. 2, no. 1, pp. 27–41, 2018.

[41] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013.

[42] I. D. Wood, J. P. McCrae, V. Andryushechkin, and P. Buitelaar, "A comparison of emotion annotation approaches for text," *Information*, vol. 9, no. 5, p. 117, 2018.