Emotion in Art-Elicited Texts, Interpreted by GPT-4o: Self vs. Third-Person Perspectives

Hiromi Narimatsu
NTT Communication Science Laboratories
Kanagawa, Japan
narimatsu@ieee.org

Keishi Nomura
Toyo University
Tokyo, Japan
nomura@toyo.jp

Xijia Wei
University College London
London, UK
xijia.wei.21@ucl.ac.uk

Nadia Bianchi-Berthouze
University College London
London, UK
nadia.berthouze@ucl.ac.uk

Shiro Kumano
NTT Communication Science Laboratories
Kanagawa, Japan
kumano@ieee.org

***Abstract*—This study investigates whether GPT-4o can serve as a proxy in emotion research by estimating human emotional states from text. Unlike prior resources such as EmoBank, which include inferred writer labels and emotionally neutral texts, we collected a dataset in which participants viewed visual artworks, reported their own emotions using the valence–arousal–dominance (VAD) model, alongside natural language descriptions. GPT-4o was prompted to estimate emotional intensity from both writer (self) and reader (third-party) perspectives, and its predictions were compared to human self-reports and reader judgments. Using pairwise comparison tasks, we evaluated agreement based on forced-choice judgments. Results show that GPT-4o aligns closely with mean human readers for valence, moderately for arousal, and poorly for dominance. Notably, for arousal, the alignment between GPT-4o's writer and reader perspectives varied depending on whether the texts shared the same emotional context—that is, whether they pertained to the same image. We also examined prompt design effects but found no benefit from intermediate reasoning for the pairwise comparison of the art-elicited texts. Our findings suggest that with appropriate task framing, LLMs like GPT-4o may support pilot studies in cognitive and affective research.**

***Index Terms*—self emotion estimation, first-person perspective, third-party perspective, VAD model, emotion from text**

I. Introduction

Emotion expression is not limited to facial expressions or gestures but is richly encoded in language. Accordingly, the task of estimating emotions from text has drawn growing attention across fields such as psychology [18], affective computing, and natural language processing (NLP) [8], [22]. With the advent of large language models (LLMs), particularly models like GPT-4o, there is increasing interest in whether such systems can interpret emotions from written language in ways that approximate human understanding [5], [32].

Most existing work in text-based emotion modeling has focused on third-party annotations, such as EmoBank [7], where ratings reflect perceived rather than experienced emotions. Notably, the EmoBank corpus [7] provides third-parties' annotations given from the first- and third-person's perspective of emotional content on the Valence–Arousal–Dominance (VAD) dimensions. The writer annotations in EmoBank are inferred by readers, not self-reported, and source texts are emotionally neutral. Consequently, alignment between expressed and perceived emotions is indirect, and the role of writer intent versus reader interpretation is difficult to evaluate empirically.

To address these limitations, we adopts an alternative design: participants first viewed visual artworks which elicit diverse emotions and allow for subjective interpretations without imposing explicit emotional labels. This approach facilitates the collection of individualized emotional responses, aligning with our goal to examine personal emotional experiences. Participants then rated their own emotions using the valence–arousal–dominance (VAD) model and described their impressions in natural language, yielding a writer-centered dataset consisting of paired self-reports and texts. We constructed two types of text pairs—based on either identical or different visual stimuli—to investigate whether shared emotional context affects the accuracy of emotional inference. Using these pairs, we evaluated agreement between human readers, GPT-4o, and the original writers under both writer- and reader-perspective prompts. In particular, we examined whether GPT-4o could accurately infer emotional states under two distinct prompting conditions: (1) estimating the writer's internal feelings (self-perspective), and (2) evaluating emotional expression from a third-party reader's perspective.

By comparing GPT-4o's outputs with both human self-assessments and third-party judgments, we aim to assess the model's ability to approximate perspective-dependent emotional reasoning. Additionally, we employ pairwise comparisons as a robust evaluation framework, enabling consistent relative judgments even for hard-to-rate dimensions like arousal and dominance [30]. Through this approach, we explore the feasibility of using LLMs not merely as analysis tools, but as pilot surrogates in human emotion research.

This study makes the following contributions:

- This study introduces a novel experimental design that links self-reported emotions elicited by visual stimuli with their associated textual descriptions, while statisti-

cally comparing conditions in which only writer-related influences operate versus those in which only content-related (i.e., targeted image) influences are present.
- It systematically compares human and GPT-4o emotion inference performance across different perspectives (i.e., first- vs. third-person perspectives), and examines how shared context (i.e., same-image vs. different-image text pairs) influences estimation consistency.
- It highlights the impact of prompt design on aligning model predictions with human judgments, offering insights into the effective use of large language models as proxies in affective and cognitive research.

II. Related Work

A. Emotion Estimation from Text

Emotion estimation from text has been widely studied in sentiment analysis, affective computing, and natural language processing. Early approaches relied on discrete models, classifying emotions into categories such as positive/negative sentiment or Ekman's basic emotions (e.g., joy, anger, sadness) [12], [22], [26]. These methods often used affective lexicons such as SentiWordNet [2], [13] or NRC Emotion Lexicon [24], and were applied mainly to short-form evaluative texts like reviews or tweets.

Recent studies have adopted dimensional models, such as Valence–Arousal–Dominance (VAD) dimensions [27], which enables richer representations of affective states along continuous scales. Datasets such as ANEW [4] and VADER [15] have supported this shift. However, most emotion estimation research remains focused on third-party annotation, without direct access to the writer's self-assessed emotional state. Consequently, the alignment between expressed and perceived emotions is often assumed rather than measured.

B. Writer vs. Reader Perspectives in Emotion Annotation

Several studies have explored how emotional meaning may differ depending on whether it is inferred by third persons but from the first- (i.e. writer's) versus third-(i.e.reader's) perspectives. A central dataset in this area is EmoBank [7], which provides VAD-based annotations from both perspectives. EmoBank distinguishes between readers' ratings perceived by their own perspective, and their ratings inferred based on assumptions about what the author likely felt when writing the text. However, EmoBank's writer labels are not self-reported but are instead approximated by annotators, and the source texts were not originally produced to convey emotional experiences. As such, the alignment between the text and actual internal states remains indirect.

Further psychological evidence supports the idea that internal emotional states are not always reliably inferred by others. Studies have shown that people overestimate the transparency of their emotions [14], [16], [19], leading to misalignments between what is felt and what is perceived. This discrepancy is particularly problematic for less observable dimensions such as arousal and dominance, which are often inferred inconsistently across raters [6]. These findings highlight a gap in current research: although many corpora include reader-oriented annotations, relatively few datasets capture self-reported emotions alongside text, especially in controlled settings where the writer intends to convey internal impressions.

C. Large Language Models and Affective Reasoning

Recent advancements in large language models (LLMs), including GPT-3 and GPT-4, have opened new possibilities for affective reasoning through text including sentiment analysis tasks, or text generation from emotional category or scores [11], [17], [21], [28], [34], [35]. These models can interpret and generate emotionally nuanced content, and some studies have begun evaluating their ability to simulate emotion understanding [3], [5]. GPT-3 has, for example, been applied to tasks involving emotion recognition and human affective judgments in dialogue [20], [36].

However, the ability of LLMs to take on specific interpretive perspectives—such as inferring emotions from the writer's point of view versus a third-party stance—remains underexplored. Tak and Gratch [31], [32] recently examined whether GPT behaves as a computational model of emotion and found that it aligns well with aggregate reader-level judgments, but does not consistently reproduce individual self-assessments associated with text.

Our study builds on this line of work by using a self-annotated dataset, where participants provide both VAD ratings and written descriptions of their emotional impressions. This allows us to evaluate whether GPT-4o can infer emotions from a writer-grounded perspective, and whether it distinguishes between self- and third-party emotional interpretations. We further explore how shared emotional context for writers, prompt design and perspective framing affect the alignment between GPT-generated outputs and human emotional reasoning.

III. Method

To verify whether GPT-4o[1] can take both writer's and reader's perspectives, we constructed an emotional dataset that captures both the emotional experiences of writers and the inferences made by readers across three affective dimensions: valence, arousal, and dominance (VAD). To elicit controlled yet naturalistic emotional responses from writers, we used emotionally evocative visual artworks as stimuli. This approach enabled us to systematically influence the emotional states being described while preserving subjective variation in textual expression. Furthermore, the use of judgment targets allowed us to perform two types of pairwise comparisons to assess whether human and GPT-4o emotional judgments are affected by writer- and target-related factors:

- **within-image pairs** (between different writers' descriptions of the same artwork)
- **within-writer pairs** (comparing descriptions of different artworks by the same writer)

[1]GPT-4o-2024-11-20 is used in this study.

These two comparisons enable both cross-perspective and within-subject analyses. Furthermore, this pairwise comparison approach based on relative judgments arguably allows more consistent and interpretable evaluations of emotional intensity [33]. Given the abstract and introspective nature of these emotional dimensions—particularly when expressed in artworks—third-party raters often face difficulties in assigning absolute numerical scores to textual expressions of emotion [29].

In the first stage, we collected data consisting of self-reported VAD ratings and free-text impression descriptions from participants (writers) who viewed visual artworks. Based on this dataset, we selected from the dataset pairs of text samples that varied either in image or writer, enabling within- and between-subject comparisons. This design also allowed us to examine whether sharing the same visual context (i.e., the same artwork) would enhance the reader's ability to infer the writer's emotional state solely from text.

In the second stage, we presented these text pairs to third-party human raters (readers) and to GPT-4o, prompting them to judge which text in each pair reflected a higher degree of valence, arousal, or dominance. This design allowed us to compare emotional reasoning across writer, reader, and model on a common evaluative framework, and to assess the consistency and divergence across perspectives.

Data collection followed the ethics review process set by the different institutions involved in the study. In addition, data were fully anonymized before being analysed.

A. Writer Data: Emotional Texts with Ratings

For collecting human writer emotional data, we focused on visual art which elicits various affective appraisals, cognitions, and reactions in the viewer. We followed the ArtEmis-EN/JP protocol [1], [25] which include emotional reaction of category selection and impression text for visual art. Participants were asked to evaluate their own impressions upon viewing each painting using a three-dimensional emotion scale consisting of valence (positive–negative), arousal (excited–calm), and dominance (dominant–submissive). Each dimension was rated on a 101-point visual analog scale (VAS), and participants also provided free-form textual descriptions of their impressions. The questions asked to the participants for each artwork are as follows:

- Please adjust the slider to indicate how positive or negative your impression was.
- Please adjust the slider to indicate how excited or calm your impression was.
- Please adjust the slider to indicate how dominant (controllable) your impression was.
- Please choose the option that best describes your impression of the image presented.
- Why and for what reason did you get that impression? Please be specific.

The paintings used in the study were a subset of 216 selected from the original ArtEmis dataset consists of 80,031 artworks, chosen to avoid skewed emotional distributions. A total of 917 participants (57.28±10.96-y (mean/SD), 54.57±10.39-y; 573M-344F), recruited from an online crowdsourcing platform, were each shown 30 paintings selected at random from a total of 216 to ensure an equal number of paintings in each category, and completed the task via a web interface. Upon completion, participants received a point-based monetary reward. Once the writers' raw VAD scores were obtained, the scores were standardized (z-scored) to eliminating individual biases in rating scale usage. This normalization was performed independently for each dimension for all writers.

To enable pairwise evaluation, we randomly selected 100 text pairs (50 within-image pairs and 50 within-writer pairs) from the created dataset. For each text pair, we assigned *which-is-higher* label for each VAD dimension (valence, arousal, dominance) based on the z-scored writer VAD scores. Specifically, a label of 1 was assigned if Sentence 1 had a higher score than Sentence 2 for the given dimension, and a label of 2 was assigned if Sentence 2 had the higher score. These binary labels were used consistently across all evaluation conditions as writers' labels. Moreover, this procedure, namely pairwise label generating based on each writer's absolute judgment on images individually, was necessary for the within-image-pair condition, in which individual judgments of two writers are to be compared. On contrary, it is not necessary for readers and GPT-4o who can give which-is-higher directly to both within-image and within-writer pairs.

B. Readers' Two-Alternative Forced Choice (2AFC) Data

For each sentence pair, we asked raters (readers) to indicate which of the two sentences in a pair of writers' texts expressed a higher level of valence, arousal, or dominance. For collecting human reader data, we recruited 10 participants — as convenience sample recruited from staff and postgraduated students from various universities. In each task, they were presented with pairs of sentences (Sentence 1 and Sentence 2) in an Excel sheet and were asked to judge which sentence reflected a higher value for each of the three VAD dimensions. Importantly, no information about the type of sentence pair (e.g., within-image or without-image) was provided to the participants, and the order of the sentence pairs was randomized to avoid order effects. To ensure consistency in interpretation, the following definitions were provided for each dimension prior to the task.

- Valence: Which sentence is more positive (positive vs. negative)?
- Arousal: Which sentence has higher arousal (excited [higher] vs. calm [lower])?
- Dominance: Which sentence expresses stronger dominance (dominant [stronger] vs. submissive [weaker])?

To reduce cognitive load and increase consistency, readers made relative comparisons rather than assigning scalar values. The same task was also presented to GPT-4o, using carefully designed prompts that instructed the model to act as a third-party reader or to infer the writer's internal state, depending on the condition.

You will be provided with a prompt and a statement describing how a specific person feels when looking at a visual art image in response to that prompt. The prompt will urge the person to explore the piece of artwork, and the statement will reflect their feelings or reactions. Based on the provided statement and its respective prompt, you are required to perform the following tasks:
- Estimate the valence rating self-reported by the person who wrote this statement, using a value between 0 (negative) and 1 (positive), with second decimal places.
- Estimate the arousal rating self-reported by the person who wrote this statement, using a value between 0 (calm) and 1 (excited), with second decimal place.
- Estimate the dominance rating self-reported by the person who wrote this statement, using a value between 0 (non-dominant) and 1 (dominant), with second decimal place.
Sentence 1: "{sentence_1}"
Sentence 2: "{sentence_2}"
Provide the output in the following JSON format:
{{
"sentence_1_valence": ...,
"sentence_2_valence": ...,
"which_is_higher_valence": ..., // 1 if sentence_1 has a higher valence, 2 if sentence_2 has a higher valence
"sentence_1_arousal": ...,
"sentence_2_arousal": ...,
"which_is_higher_arousal: ..., // 1 if sentence_1 has a higher arousal, 2 if sentence_2 has a higher arousal "sentence_1_dominance": ...,
"sentence_2_dominance": ...,
"which_is_higher_dominance": ..., // 1 if sentence_1 has a higher dominance, 2 if sentence_2 has a higher dominance
}}

Fig. 1: Prompt used for GPT-writer.

Two texts will be shown to you. Please evaluate them according to the tasks below:
- Valence: Which sentence is more positive (positive vs. negative)?
- Arousal: Which sentence has higher arousal (excited [higher] vs. calm [lower])?
- Dominance: Which sentence expresses stronger dominance (dominant [stronger] vs. submissive [weaker])?
Sentence 1: "{sentence_1}"
Sentence 2: "{sentence_2}"
Provide the output in the following JSON format:
{{
"which_is_higher_valence": ..., // 1 if sentence_1 has a higher valence, 2 if sentence_2 has a higher valence
"which_is_higher_arousal": ..., // 1 if sentence_1 has a higher arousal, 2 if sentence_2 has a higher arousal
"which_is_higher_dominance": ..., // 1 if sentence_1 has a higher dominance, 2 if sentence_2 has a higher dominance
}}

Fig. 2: Prompt used for GPT-reader.

This setup allows for a direct evaluation of agreement across the following sources: (1) Human writers' self-reported VAD ratings, (2) Third-party reader judgments (pairwise decisions), (3) GPT-4o predictions under two prompting conditions: writer's self-perspective and reader-perspective.

C. GPT-4o's Two-Alternative Forced Choice Data

For each pair of sentences, GPT-4o was presented with both descriptions and asked to select which one reflected a higher degree of a given emotion dimension. Two prompting conditions were used:

- GPT writer (writer-perspective prompt): GPT was instructed to estimate the emotions experienced by the person who wrote the given text, using the prompt shown in Fig. 1. The model was asked to estimate the emotional intensity experienced by the original writer of each sentence across the three VAD dimensions. For each sentence in the pair, GPT was instructed to generate VAD ratings, and then determine which of the two texts reflected a higher value along the target dimension. This procedure was designed to mirror the format of the human writer's self-assessment task, in which writers had directly assigned numerical VAD values to their own emotional impressions.
- GPT reader (reader-perspective prompt): The model was asked to act as an external reader and compare the two sentences to judge which one conveyed a stronger emotional intensity in each VAD dimension. Unlike the writer condition, no intermediate score generation was required; the model directly selected the more intense text based on the perception of an outside observer. The prompt design was aligned with the instructions given to human third-party raters to ensure consistency. Fig. 2 is the used prompt.

To eliminate randomness unrelated to the intended perspective manipulation, the temperature parameter was set to 0 in all GPT-4o generations. The outputs obtained under these two conditions are hereafter referred to as GPT-writer and GPT-reader, respectively. The resulting judgments were then compared against human writer labels and third-party reader decisions to assess the model's perspective-taking ability and alignment with human emotional reasoning.

IV. Evaluation

We first evaluate the agreement in responses across writer–reader pairs for both human (HMN) and GPT settings, focusing on the differences introduced by the respective perspectives. Furthermore, to carefully investigate how perspective shifts introduced by prompts and variations in inference methods affect the responses, we additionally assess agreement under two specific conditions: (1) pairs sharing the same contextual basis (within-image), and (2) pairs differing in context but produced by the same writer (within-writer). We also evaluate the consistency of responses when VAD scores are estimated prior to the *which is higher* judgment to evaluate the stability of GPT's answers.

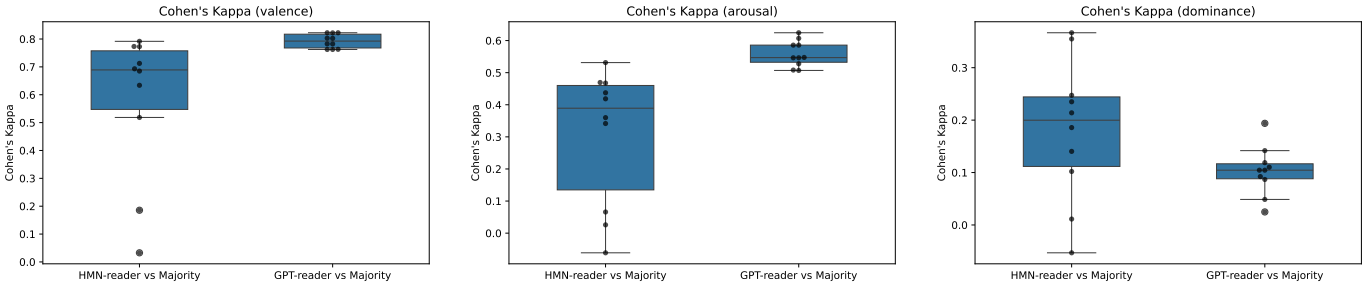
A. Writers vs. readers: GPT-4o vs. human perspectives alignment

To compare the outputs of each condition, we assessed inter-rater agreement using Cohen's kappa [9], which is identical to Krippendorff's alpha (used in previous studes, such as [32]) for binary variables. Pairwise comparisons were conducted for all combinations among the four conditions: HMN-writer, HMN-reader, GPT-writer, and GPT-reader, where HMN represents real human. Furthermore, McNemar's test [23] was employed to evaluate the statistical significance of differences between each pair of methods.

Table I shows the overall agreement of all condition pairs. For the HMN-reader condition, we used a majority vote from

TABLE I: Cohen's κ and McNemar's test statistics (χ^2, p) across comparison pairs and emotion dimensions.

Pair	Valence			Arousal			Dominance		
	κ	χ^2	p	κ	χ^2	p	κ	χ^2	p
HMN writer vs. HMN reader	0.304	1.333	0.248	-0.057	2.250	0.134	-0.039	0.500	0.480
GPT writer vs. GPT reader	0.786	0.100	0.752	0.593	1.389	0.239	0.568	8.471	0.004
HMN writer vs. GPT writer	0.322	1.333	0.248	0.056	1.333	0.248	0.055	2.250	0.134
HMN reader vs. GPT reader	0.850	0.571	0.450	0.635	0.062	0.803	0.151	12.250	0.000
HMN writer vs. GPT reader	0.360	1.333	0.248	-0.077	3.200	0.074	0.076	5.143	0.023
HMN reader vs. GPT writer	0.850	0.571	0.450	0.548	1.250	0.264	0.066	1.730	0.188



(a) Valence (b) Arousal (c) Dominance

Fig. 3: Cohen's κ comparison between (human vs. majority) vs. (GPT-4o vs. majority). Each plot represents a comparison between the majority vote excluding a single respondent and either (a) the excluded respondent or (b) GPT-4o. Since there are 10 respondents, a total of 10 plots are presented.

10 human participants to derive consensus labels. This enabled direct comparison between human and GPT outputs under controlled conditions. The results showed a clear gap between human writer and reader perspectives, with substantial disagreement—consistent with prior findings [6]. In contrast, GPT showed relatively little difference between its writer and reader prompts, suggesting that it may not fully capture the cognitive shift between perspectives. When comparing GPT to humans under the same perspective, GPT aligned more closely with human reader judgments than with writer self-assessments. This pattern is in line with previous work using datasets like EmoBank [7], where third-person annotations proved more consistent and predictable than self-reports. Additionally, HMN-writer vs. GPT-reader and HMN-reader vs. GPT-writer comparisons supported this trend. These cross-perspective results confirm that GPT mirrors reader reasoning more reliably than a writer's internal emotional states.

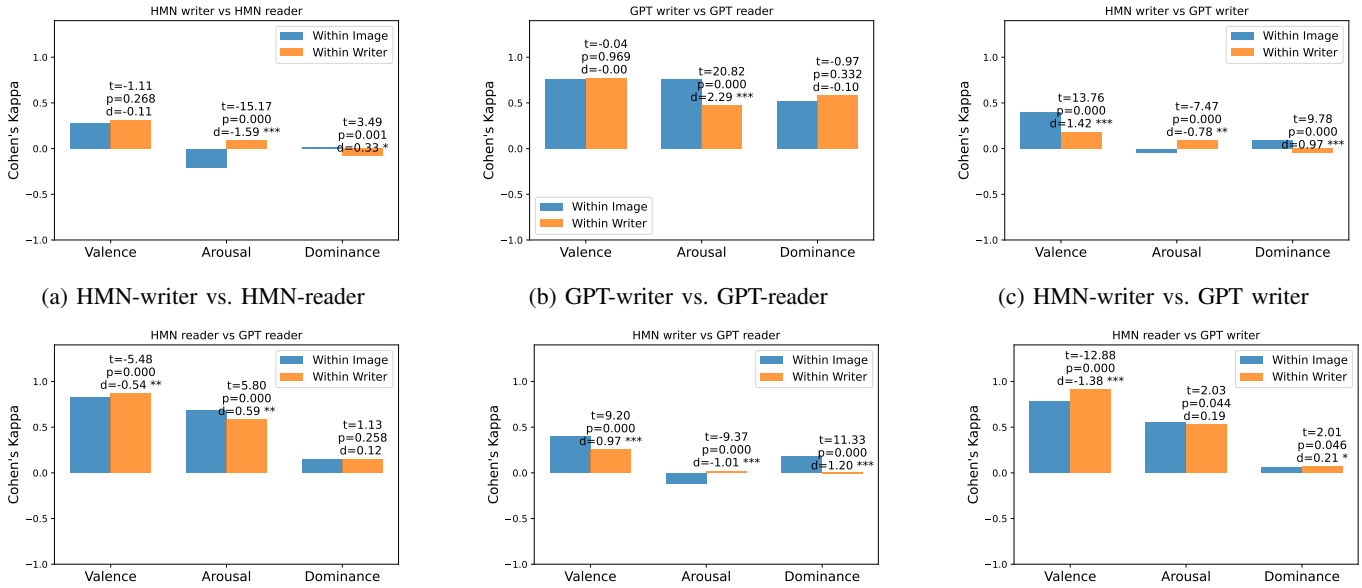
Three VAD dimensions have different tendency. In human writer vs. reader comparisons, valence showed moderate agreement, while arousal and dominance showed much lower consistency. This supports prior research [10], [30] noting the inherent difficulty in judging arousal and dominance compared to valence. A similar trend was observed in the agreement between HMN-reader and GPT-reader, with Cohen's κ values of 0.850 for valence, 0.635 for arousal, and 0.151 for dominance. These results suggest that GPT-4o effectively captures valence and, to some extent, arousal from a reader's perspective. However, dominance remains challenging, highlighting limitations in both human interpretation and model inference in this dimension.

To further explore alignment between individual human readers and consensus judgments, we calculated the agreement between each human reader responses and the human reader majority, excluding the reader from the majority calculation. Similarly, we compared GPT-reader perspective outputs against a rotating majority of human readers, each time excluding one participant. The results are shown in Fig. 3.

The results reveal two notable patterns. First, inter-human agreement (a) shows relatively high consistency in the valence dimension across readers (mean $\kappa \approx 0.72$), indicating shared interpretive norms for positivity/negativity. In contrast, arousal and especially dominance show substantial variability among participants. Some readers display near-zero or even negative κ values for dominance, suggesting low consensus and possibly differing internal definitions or thresholds for that dimension.

Second, GPT-reader agreement with the human majority mirrors these trends in dimensional difficulty. Valence again shows high agreement (mean $\kappa \approx 0.80$), indicating that GPT-4o aligns closely with human consensus in assessing emotional positivity. Arousal yields moderate agreement (mean $\kappa \approx 0.55$), while dominance remains low (mean $\kappa \approx 0.10$), consistent with earlier pairwise evaluation results.

Interestingly, GPT-4o's agreement with the human majority in valence is often comparable to or even higher than that of some individual human raters. This suggests that, in certain dimensions, the model may serve as a stable proxy for aggregated human judgments. However, the consistently low agreement across both humans and GPT highlights the dominance



(a) HMN-writer vs. HMN-reader (b) GPT-writer vs. GPT-reader (c) HMN-writer vs. GPT writer

(d) HMN-reader vs. GPT-reader (e) HMN-writer vs. GPT-reader (f) HMN-reader vs. GPT-writer

Fig. 4: Cohen's κ coefficients comparing agreement levels on valence, arousal, and dominance across different writer–reader pairings. Two conditions are contrasted: within-image (same image / different writer) and within-writer (different image / same writer). Each subplot represents a specific combination of human (HMN) and/or GPT-based writer and reader, illustrating the consistency of emotional dimension judgments under each condition. Statistical significance is calculated using t-tests and calculated effect sizes using Cohen's d.

dimension as particularly ambiguous and difficult to model, reinforcing its known complexity in affective annotation.

These findings emphasize the need to treat emotional dimensions differently when evaluating model performance and suggest that LLMs like GPT-4o may best function as proxies for consensus-level interpretations, particularly in affectively simpler dimensions like valence and to certain extent arousal.

B. Within-image vs. within-writer level of agreement

Fig. 4 presents a comparison of inter-rater agreement across different combinations of human and GPT-based writer/reader pairings, using Cohen's κ. Two conditions were contrasted: within-image (same image, different writer) and within-writer (same writer, different image). Across all pairings, valence consistently shows the highest agreement, particularly in human–human and human–GPT reader comparisons (Fig. 4a,d,f). Notably, GPT-reader estimates align closely with both human writers and readers, especially under the within-image condition, suggesting that GPT-4o can robustly generalize across individuals when inferring valence.

For arousal, agreement was moderate and more sensitive to rater combination and condition. Human–GPT comparisons (e.g., 5d, 5f) showed relatively stable but lower κ values than valence, while HMN-writer vs HMN-reader comparisons (5a) revealed greater variability between within-image and within-writer conditions. This indicates that arousal is less consistently interpreted across perspectives and that intra-personal variability may also be high.

In contrast, dominance yielded low agreement across all comparisons, with κ values near zero or even negative in some pairings. This pattern reaffirms the previous finding that dominance is difficult to assess reliably from textual impressions, for both humans and GPT-4o, regardless of perspective or pairing condition.

Together, these results highlight that GPT-4o can approximate human-like emotion reasoning for valence, and to some extent arousal, but struggles with dominance, consistent with prior analyses. Moreover, the within-image condition tends to yield higher agreement than within-writer, suggesting that emotional interpretation is more stable across shared perceptual inputs than across internal state variation.

TABLE II: Comparison of GPT-reader responses under two conditions: with and without prior estimation of VAD (Valence, Arousal, Dominance) scores. The table reports the mismatch ratio between the two conditions, along with the average difference in estimated scores between Sentence1 and Sentence2 for both answer-matched and answer-differed cases.

Dimension	**Mismatch Ratio**	**Avg. Score Difference**	
		Answer-matched	*Answers-differed*
Valence	0.09	0.213	0.067
Arousal	0.15	0.210	0.090
Dominance	0.18	0.160	0.117

TABLE III: Cohen's κ comparison between HMN-reader majority answer vs. two GPT-reader conditions (1) with prior estimation of VAD scores, and (2) without prior estimation of VAD scores before answering the *which is higher* question.

Dimension	With vs. Without	HMN vs. With	HMN vs. Without
valence	0.851	0.786	0.850
arousal	0.616	0.616	0.635
dominance	0.546	0.086	0.151

C. Impact of intermediate reasoning

To investigate the effect of prior VAD score estimation on pairwise emotion judgments, we compared two conditions: (1) a two-step approach in which GPT-4o first estimated VAD scores for each text and then answered the *which is higher* question based on those scores, and (2) a direct comparison approach in which the model answered the *which is higher* question without estimating individual scores beforehand (Table. II,III). The results showed that response changes tend to occur when the difference in predicted scores between the two sentences is small, and no significant difference in performance between the two conditions, as confirmed by McNemar's test (valence: $\chi^2 = 0.125$, $p = 0.724$; arousal: $\chi^2 = 1.25$, $p = 0.264$; dominance: $\chi^2 = 7.68$, $p = 0.0056$). Although the dominance dimension exhibited a statistically significant difference, this likely reflects the low inter-annotator agreement commonly observed in human annotations of dominance. Given the generally low reliability of dominance responses, no meaningful conclusions can be drawn from this result. On the contrary, agreement with the human majority was slightly higher in *without score estimation* condition. This may be because humans typically make such comparative judgments without assigning explicit numerical scores, and the latter approach may better align with that natural decision-making process. This suggests that GPT-4o's pairwise emotion judgments are not necessarily enhanced by explicitly generating intermediate VAD scores, and that direct comparative reasoning may be equally effective in this context.

V. Discussion

We examined whether GPT-4o can infer human emotional states from text in ways that approximate either self-assessed or third-party interpretations. Our findings consistently showed that GPT-4o aligns more closely with third-party judgments, particularly for valence, while its agreement with writer self-assessments was significantly lower—especially in the more cognitively abstract dimensions of arousal and dominance.

These findings align with recent work by Tak and Gratch [32], showed that GPT-4 emulates *average third-person emotional reasoning*, but fails to refrect self-reported emotions. They found that the model defaults to socially shared norms, even when prompted to take a first-person perspective. In contrast, our study uniquely examined whether shared emotional context—specifically, describing the same artwork —affects the accuracy of emotion inference. By controlling for stimulus and varying perspective, our design isolated the effect of contextual alignment, revealing that shared context improves agreement, particularly in arousal estimation.

Prompt design for GPT-4o requires careful attention, especially when trying to model unique interpretations. In a preliminary experiment, we tested a modified version of the GPT-reader prompt in which only the perspective-related phrasing was changed—substituting the GPT-writer prompt's wording into the reader context. However, this minor adjustment did not impact the model's output at all. The generated responses were the same as those from the GPT-writer condition. This finding shows that GPT-4o doesn't meaningfully differentiate between prompts unless the shift in perspective is clearly operationalized through task-aligned instructions. This emphasizes the importance of prompt fidelity-not just in terms of perspective labels, but also in aligning the model's role and cognitive framing with that of the human evaluators.

While these limitations reflect the current capabilities of GPT-4o, they may not generalize to future models. As LLMs continue to evolve—with newer iterations like GPT-4.5 reportedly improving in contextual reasoning and perspective tracking—it is plausible that upcoming systems will better distinguish between self- and other-perspectives in emotion understanding. Enhancing such interpretive flexibility could make LLMs valuable not only for third-party inference but also for capturing the *subjective, writer-centered dimensions* of emotion. Future work should test whether next-generation models can reduce these alignment gaps, especially in dimensions like dominance where current performance is low.

VI. Conclusion

This study explored whether GPT-4o can serve as a proxy for human emotion estimation by comparing its inferences against human self-reported and third-party judgments across the VAD dimensions. Using text produced in response to visual stimuli, we systematically evaluated agreement under both writer- and reader-perspective prompts and different emotional text conditions. The results demonstrate that GPT-4o shows strong alignment with third-party reader judgments—particularly for valence—but performs less reliably in modeling self-reported emotional states. These limitations were most apparent in dimensions like dominance, which involve more subtle, relational cues. GPT-4o was also found to be affected by whether the text shares the same emotional context.

Taken together, these results suggest that GPT-4o can serve as a useful pilot tool in emotion research, especially for capturing broadly shared interpretations. Still, its limitations in perspective sensitivity and complex dimensions like dominance highlight the need for careful prompt design and dimension-specific consideration when applying LLMs in human-centered studies.

Ethical Impact Statement

Estimating the target person's emotion represented in self-reported text finds application in various fields. Given the considerable impact that LLM applications can have on user emotional well-being, issues of transparency and fairness are

of special importance, reflecting current discussions within the affective computing community. This paper contributes to bridging the gap between affective science and technological implementation by providing new insights into the accuracy of emotional state estimation in state-of-the-art technologies capable of emotion-aware interaction, as well as the discrepancies that may arise between model predictions and individuals' self-reported experiences. While emotion estimation from text holds significant potential, it also carries important risks, including the reinforcement of biases and stereotypes derived from training data, violations of individual privacy, and the potential for emotional misinterpretation or misclassification.

Ethical considerations are needed to address these potential issues.

REFERENCES

[1] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas. Artemis: Affective language for visual art. In Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 11569–11579, 2021.

[2] S. Baccianella, A. Esuli, F. Sebastiani, et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec, volume 10, pages 2200–2204. Valletta, 2010.

[3] M. Binz and E. Schulz. Using cognitive psychology to understand gpt-3. Proceedings of the National Academy of Sciences, 120(6):e2218523120, 2023.

[4] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ⋯, 1999.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[6] 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, pages 1–12, 2017.

[7] S. Buechel and U. Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. arXiv preprint arXiv:2205.01996, 2022.

[8] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on affective computing, 1(1):18–37, 2010.

[9] J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46, 1960.

[10] L. De Bruyne, O. De Clercq, and V. Hoste. Annotating affective dimensions in user-generated content: Comparing the reliability of best–worst scaling, pairwise comparison and rating scales for annotating valence, arousal and dominance. Language Resources and Evaluation, pages 1–29, 2021.

[11] I. N. Debess, A. Simonsen, and H. Einarsson. Good or bad news? exploring gpt-4 for sentiment analysis for faroese on a public news corpora. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 7814–7824, 2024.

[12] P. Ekman. An argument for basic emotions. Cognition & emotion, 6(3-4):169–200, 1992.

[13] A. Esuli, F. Sebastiani, et al. Sentiwordnet: A publicly available lexical resource for opinion mining. In LREC, volume 6, pages 417–422, 2006.

[14] T. Gilovich, K. Savitsky, and V. H. Medvec. The illusion of transparency: biased assessments of others' ability to read one's emotional states. Journal of personality and social psychology, 75(2):332, 1998.

[15] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media, volume 8, pages 216–225, 2014.

[16] B. Keysar and A. S. Henly. Speakers' overestimation of their effectiveness. Psychological Science, 13(3):207–212, 2002.

[17] 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.

[18] O. Kjell, K. Kjell, D. Garcia, and S. Sikström. Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. Psychological Methods, 2018.

[19] J. Kruger, N. Epley, J. Parker, and Z.-W. Ng. Egocentrism over e-mail: Can we communicate as well as we think? Journal of personality and social psychology, 89(6):925, 2005.

[20] M. Lammerse, S. Z. Hassan, S. S. Sabet, M. A. Riegler, and P. Halvorsen. Human vs. gpt-3: The challenges of extracting emotions from child responses. In 2022 14th International conference on quality of multimedia experience (QoMEX), pages 1–4. IEEE, 2022.

[21] F. Lecourt, M. Croitoru, and K. Todorov. " only chatgpt gets me": An empirical analysis of gpt versus other large language models for emotion detection in text. arXiv preprint arXiv:2503.04831, 2025.

[22] B. Liu. Sentiment analysis and opinion mining. Springer Nature, 2022.

[23] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2):153–157, 1947.

[24] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. Computational intelligence, 29(3):436–465, 2013.

[25] H. Narimatsu, R. Ueda, and S. Kumano. Cross-linguistic study on affective impression and language for visual art using neural speaker. In 10th Int'l Conf. on Affective Computing and Intelligent Interaction, pages 1–8, 2022.

[26] B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval, 2(1–2):1–135, 2008.

[27] J. A. Russell. A circumplex model of affect. Journal of personality and social psychology, 39(6):1161, 1980.

[28] K. Schaaff, C. Reinig, and T. Schlippe. Exploring chatgpt ' s empathic abilities. In 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–8, 2023.

[29] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9):e73791, 2013.

[30] I. Siegert, R. Böck, and A. Wendemuth. Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. Journal on Multimodal User Interfaces, 8:17–28, 2014.

[31] 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), pages 1–8. IEEE, 2023.

[32] A. N. Tak and J. Gratch. Gpt-4 emulates average-human emotional cognition from a third-person perspective. arXiv preprint arXiv:2408.13718, 2024.

[33] L. L. Thurstone. A law of comparative judgement. Psychological Review, 34:278–286, 1927.

[34] R. Ueda, H. Narimatsu, Y. Miyao, and S. Kumano. Emotion-controllable impression utterance generation for visual art. In 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–8. IEEE, 2023.

[35] R. Ueda, H. Narimatsu, Y. Miyao, and S. Kumano. Vad emotion control in visual art captioning via disentangled multimodal representation. In 12th Int'l Conf. on Affective Computing and Intelligent Interaction, 2024.

[36] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin. Is chatgpt equipped with emotional dialogue capabilities? arXiv preprint arXiv:2304.09582, 2023.