RESEARCH ARTICLES

The Comparative Emotional Capabilities of Five Popular Large Language Models

Nathan Klapach[a]

Critical Debates in Humanities, Science and Global Justice
Vol. 2, Issue 1, 2024

Large language models (LLMs) are artificial intelligence algorithms that apply deep learning to large data sets to understand, summarize, generate, and predict content. The most common application of LLMs is generative AI for text-based content. Public awareness of LLMs has greatly increased with the recent release of ChatGPT3. While LLMs are powerful conveyors of information, this study explores their emotional capabilities. By examining 5 popular LLMs (BingAI, ChatGPT3.5, GoogleBard, ChatGPT4, HuggingChat) and their ability to comprehend, mimic, and convey emotion, this study attempts to determine which of these LLMs can best process the five core emotions of joy, sadness, anger, fear, and disgust. Based on the results of the study's survey, the most effective LLM with respect to emotional content was ChatGPT4, followed by ChatGPT3.5, HuggingChat, GoogleBard, and BingAI. A greater understanding of the comparative emotional capabilities of LLMs will be pivotal to assessing and predicting their potential for therapeutic, medical, natural language processing, and personal use.

Author Summary

In recent years, artificial intelligence has improved dramatically, as Large Language Model (LLM) AIs show an unprecedented ability to process massive amounts of information. These LLMs have become effective interpreters of human emotion, taking roles as medical counselors and therapy aids. In this research paper, I explore which of the five most popular LLMs (BingAI, ChatGPT3.5, GoogleBard, ChatGPT4, HuggingChat) are the most emotionally saavy. Using human judges to evaluate each LLM's capability to understand, mimic, and generate emotional content, I determined that ChatGPT4 was most emotionally capable followed by ChatGPT3.5, HuggingChat, GoogleBard, and BingAI. Emotional intelligence is a key element of what it means to interact with a human in high-level discourse. By understanding which LLM is the most emotionally capable, we can more accurately assess and predict each model's potential for therapeutic, medical, linguistics, and personal use.

a Nathan is a high school student at the Davidson Academy,

Introduction

Artificial Intelligence (AI) has captivated humanity since its theoretical conception more than a century ago. For most of this time, the idea of a non-human entity communicating at or above the level of humans has largely been relegated to the realm of science fiction. However, the public release of Chat-GPT 3 in 2022 spurred a boom in AI awareness and interest, and human-machine communication has become much more commonplace.

Mainstream chatbot LLMs have shown a remarkable ability to comprehend, structure, and deliver information. For example, ChatGPT4 boasts some remarkable academic feats, such as scoring in the top 10% on the attorney bar exam (OpenAI, 2023) and the SAT (OpenAI, 2023). But while precise displays of factual information are part of communication, emotional connection permeates the vast majority of human interactions (Cheshin et al., 2011; Prochazkova & Kret, 2017). As a result, the future of LLM development will rely on its ability to comprehend and communicate with humans on an emotional level (Merriam, 2022). The purpose of this study is to determine the most emotionally mature LLM based on its ability to understand, mimic, and generate emotional content relating to joy, sadness, anger, fear, and disgust.

Methods

Literature Review

Communication with humans on an emotional level requires some degree of connection to human emotions (Prochazkova & Kret, 2017). This is a given for human-to-human communication, as people experience a similar set of emotions. Since the early 20th century, psychologists have attempted to condense every human emotion into "core emotions." Psychologist Paul Ekman concluded that the core emotions are anger, surprise, disgust, joy, fear, and sadness (Ekman, 1991). Recent studies have challenged this list, with the University of Glasgow finding that surprise/fear and anger/disgust occupy the same core emotion; this study condenses the list of core emotions into anger, joy, fear, and sadness (Jack et al., 2014). Given the debate over whether disgust and anger are the same emotion (Kollareth et al., 2022; Russell & Giner-Sorolla, 2011, 2013), the five core emotions of anger, joy, fear, sadness, and disgust will be used for the purposes of this study.

Measuring emotions is incredibly difficult. There is no way to genuinely ascertain whether an emotion is being felt or just expressed (Grose, 2011). While static emotion (i.e. what someone is feeling at any given moment) is difficult to ascertain, emotional development over time is far easier to measure. Examining patterns in emotional development reveals a spectrum of proto-emotional intelligence ranging from genuine human levels of emotion and the indifference of a machine or infant (Bridges, 1932; Churchill & Lipman, 2016). There are three milestones in human emotional development: (1) sign-mediated emotion systems from ages 0-2 years old (comprehension); (2) intrapersonal regulation from ages 2-5 (mimicry); and

(3) internalization of emotional expression at ages 5+ (generation) (Holodynski, 2008). Together, these three stages most accurately depict human emotional growth. While far from human, a LLM approximation of human emotion would need to perform adequately in each of these three stages (Ho, 2022).

Existing literature has been fairly unified in concluding that AI has the capacity to comprehend or understand emotional prompts. In the medical sector, AI is being used to interpret diagnoses, which requires a high degree of emotional comprehension (Alloghani et al., 2022). LLMs have been particularly useful, as chat-bots have connected with patients on an emotional level to help them overcome stressors (Jo et al., 2023; Nasiri, 2023; Zheng et al., 2022). Within the sphere of social media, AI language models have also been successful in interpreting the underlying emotional cues of messages posted to the platform. Using sentiment detection algorithms, language models can comprehend emotions with pinpoint accuracy (AlBadani et al., 2022; Moshkin et al., 2022) across different languages and cultures (Alhabari et. al. 2021). Existing literature demonstrates that AI has the potential for high-level emotional comprehension, and that current LLMs can accurately understand emotions (Frey et al., 2019; Heaton & Schwartz, 2020).

The other two elements of emotional maturity, mimicry and generation, are also well supported by current literature. Pre-trained language models have been incredibly efficient at paraphrasing and mimicking emotional stimuli (Casas et al., 2021), a phenomenon that has recently been found in LLMs (Serapio-García et al., 2023). In fact, LLM mimicry of emotions is so powerful that it often leads the AI to internalize political and moral beliefs (Simmons, 2023). LLM generation of emotional content demonstrates similar competence and promise. LLMs are highly adept at creating emotional content (Santhanam & Shaikh, 2019), showing the ability to adapt intensity and mood (Goswamy et al., 2020). As a result, LLMs have shown above average levels of empathy (Patel & Fan, 2023) and high EQ scores (Wang et al., 2023).

Recent literature has dispelled the notion that LLMs are consigned to robotic, emotionless communication; instead, research shows these models demonstrate remarkable emotional maturity. Unfortunately, not all LLMs are created equally, with different models showing different capacities for informational and emotional communication. While there is one comparative analysis examining the EQ of different LLMs (Wang et al., 2023), qualitative analyses offer a unique perspective into AI emotional ability (Casas et al., 2021). As a result, the purpose of this study is to conduct a qualitative, human-judged analysis into the emotional capabilities of 5 mainstream Chatbot Language-Model AIs. A prompt-based analysis will be used given that it is the most effective way to gauge LLM efficiency (Li et al., 2023; Rosenfeld, 2021; Rosenfeld & Richardson, 2019).

Experimentation

The experimentation will be completed in 3 different studies, which are then combined to create a final emotional proficiency score out of 1000. A higher proficiency score will mean that an LLM is more capable of emotional behavior and is calculated by the sum of 3 separate understanding, mimicry, and generation scores.

Participants and ethics

Collected data was qualitatively judged by 12 anonymous respondents. These judges were randomly selected and represent different cultural, economic, education, racial, and religious backgrounds. Before presenting each judge with a survey, participants were given a brief instruction manual to ensure score consistency. Each survey was conducted through a Google Form and anonymous responses ensured that no contact or personal information could be leaked. No judges under 18 years old were employed for this experiment and all data is recorded with written consent of each participant.

Experiment 1: Comprehension

For the first experiment, each LLM will be presented with 5 different flash-fiction stories designed to encapsulate the emotions of anger, sadness, joy, disgust, or fear. These stories were selected from a pool of 15 total stories from an online flash-fiction forum, with 5 independent judges selecting a story for each emotion. This was done to ensure the emotional undertones of each story are equally understandable, literal, and deep. Each LLM was then prompted to identify the emotions present in each story and instructed to return a brief explanation. Between each prompt, the LLM was closed and rebooted.

The responses of the LLM were then graded out of 300 points with LLM names and responses being kept wholly separate. This score represents the cumulative subscores of each story, with each story response scored out of 60. Each story was scored based on its comprehensibility, accuracy, and depth of the emotional response.

Comprehension Experiment grading criteria were as follows: (1) Comprehensibility: how understandable and human is the response (20 points); (2) Accuracy: how similar the generated emotional description is to the same story (20 points); (3) Depth: how descriptive was the emotional response (20 points). Each criteria was graded out of 20 points with 20 being perfectly comprehensible, accurate, and deep and 0 being completely incomprehensible, inaccurate, and shallow.

Experiment 2: Mimicry

For the second experiment, mimicry, the same set of 5 stories was once again utilized. Each LLM was prompted with each of the 5 emotional stories and instructed to construct 5 stories of similar tone, style, and emotional impact to those given previously. Between each prompt, the LLM was once again closed and rebooted.

The responses were then graded by the same panel of 12 judges out of 300 points. Like the previous experiment, this total represents the sum of each story (60 points) which in turn is made up of 20 points each for similar tone, emotional accuracy, and emotional depth. Similarity of tone represents how close the voice and message are to the original story.

Mimicry grading criteria were as follows: (1) Similarity of Tone: how close the voice and message are to the original story (20 points); (2) Accuracy: how closely the emotions of the original story were replicated (20 points); and (3) Depth: how descriptive was the emotional response (20 points).

Experiment 3: Generation

For the third and final experiment, each LLM was asked to create a script-like conversation in which five family members have an argument about which restaurant to eat out at that night. This scene was selected because it required banal yet poignant portrayal of emotions in a distinctly human and relatable experience. Since the purpose of this experiment is to determine how effectively LLMs can relate on an emotional level with humans, this prompt is fitting. With each member of the family representing one core emotion (the mother being joy, father being sadness, oldest brother being fear, middle sister being anger, and youngest brother being disgust), the LLM was then instructed to write a script within the parameters of the prompted setting and characters.

Then, each 500 word script was graded out of 400 points by the same panel of 12 judges. Generation is scored 100 points more than mimicry or comprehension because it is the highest form of emotional interaction (Holodynski, 2008). Unlike the previous experiments, generation is judged out of 4 criteria each worth up to 100 points: depth, accuracy, realism, and impact.

Generation Experiment grading criteria were as follows: (1) Depth: the emotional complexity of the characters, story, and dialogue (100 points); (2) Accuracy: how well each character represents their assigned emotions (100 points); (3) Realism: how realistic the emotional portrayal was to human behavior (100 points); and (4) Impact: the emotional power of the script meant to account for the subjectivity of human emotional comprehension (100 points).

Results

Comprehension Experiment

Due to the length of the experiment and the difficulty in finding participants willing to contribute the time necessary to complete it, the exact Comprehensibility, Accuracy, and Depth values were not measured during the Comprehension experiment. Rather, the composite scores for each emotion were recorded and their values are expressed in the below table.

Table 1. Comprehension Experiment Results

	Bing AI	ChatGPT3.5	Google Bard	ChatGPT4	Hugging Chat
Joy (60)	32.6	53.6	45.8	53.4	43.0
Sadness (60)	15.2	41.8	47.0	56.2	50.0
Fear (60)	30.4	52.4	44.2	52.6	48.6
Anger (60)	29.8	51.6	50.0	53.4	49.0
Disgust (60)	27.0	51.8	42.2	55.8	51.0
Total (300)	135.0	252.0	229.0	271.4	241.6

Mimicry Experiment

Like the Comprehension Experiment, the Mimicry Experiment was unable to measure the exact Tone, Accuracy, and Depth values. Rather, the composite scores for each emotion were recorded and their values are expressed in the below table.

Table 2. Mimicry Experiment Results

	Bing AI	ChatGPT3.5	Google Bard	ChatGPT4	Hugging Chat
Joy (60)	35.6	50.8	29.2	48.0	32.0
Sadness (60)	29.6	57.0	39.8	49.8	45.6
Fear (60)	31.2	56.2	43.4	53.2	28.0
Anger (60)	24.0	49.6	28.8	49.8	24.6
Disgust (60)	20.4	47.6	41.8	55.6	10.8
Total (300)	140.8	261.2	193.0	256.3	141.0

Generation Experiment

Like the Comprehension and Mimicry experiments, an individual breakdown of the Depth, Accuracy, Realism, and Impact is not available. Rather than adding the totality of each emotional value together to create a composite score, the integration of all 5 emotions into one story meant the script was evaluated singularly. The results of the Generation Experiment are present below.

The scores of the Generation Experiment were as follows: (1) Bing AI: 152.0; (2) ChatGPT3.5: 347.5; (3) Google Bard: 255.3; (4) ChatGPT4: 361.8; and (5) HuggingChat: 336.7.

Total Results

The composite results of each experiment are expressed in the chart below. Ultimately, ChatGPT4 places highest with a score of 889.5 followed by ChatGPT3.5 with 860.7, HuggingChat with 719.3, GoogleBard with 677.3, and BingAI with 427.8.

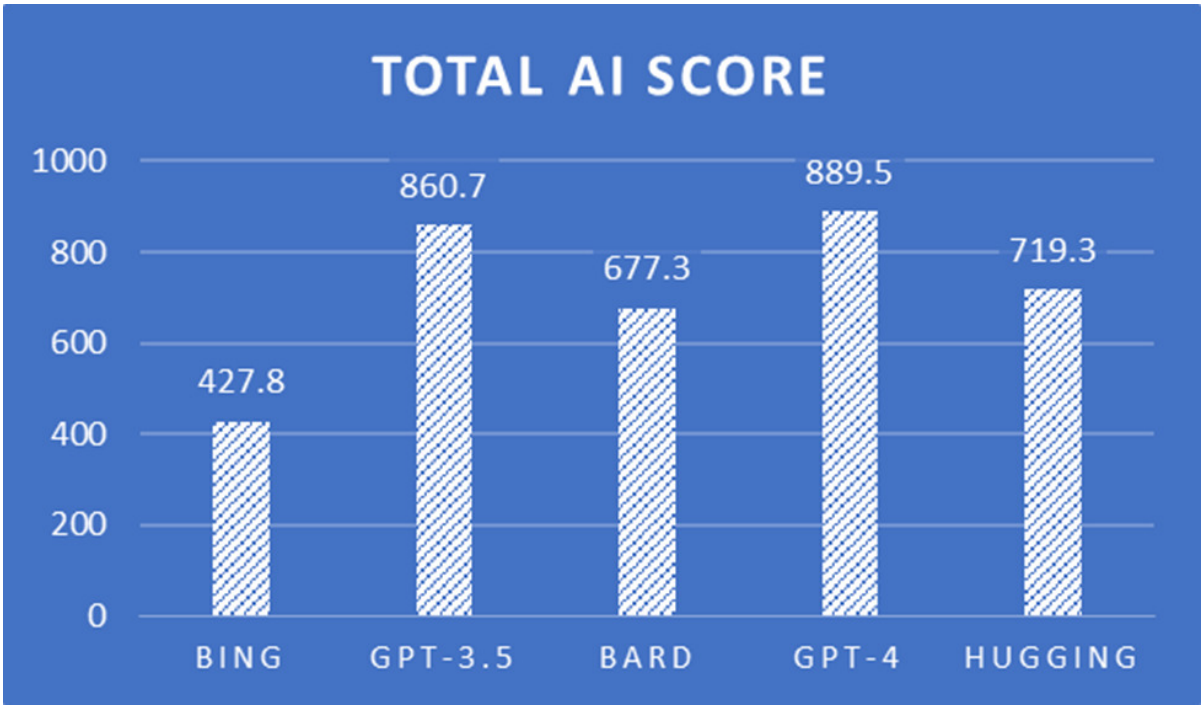

Figure 1. Total score of each LLM

Additional Findings

Comprehension Experiment results were fairly consistent. The participants tended to give even rankings within one standard deviation of the mean. No LLM varied drastically from emotion to emotion, with each emotion existing in a range. No ChatGPT4 emotion was less than the greatest Bing AI emotion. The ranking for Comprehension Experiment from first to last is as follows: ChatGPT4, GPT3.5, Hugging Chat, Google Bard, Bing AI.

Mimicry Experiment results had far greater variability in two respects. First, individual rankings were far more inconsistent and there were significant outliers. Two respondents were in complete disagreement on each story's score, each leaving highly conservative or liberal scoring more than two standard deviations from the mean. One participant rated a story a *56/60* while the other simply left a *1*. This trend continued seemingly randomly for each story, so both participants were dropped from the data as significant outliers. The effective sample size for the Mimicry Experiment is 10 rather than 12. Peculiarly, these two respondents shared no such disagreement for the other two experiments. Second, emotional rankings per AI were far more inconsistent. For example, Hugging Chat scored an above average 45.6/60 in sadness, but a record low 10.8/60 in disgust.

Discussion

Analysis of Performance across LLMs

GPT4 emerged as the superior model, achieving the highest total score of 889.5/1000, followed by GPT3.5 with a total score of 860.7/1000. This suggests that advancements and refinements in model architecture, as reflected in the progression from GPT3.5 to ChatGPT4, correlate with improved emotional responsiveness and acuity, mirroring a closer alignment with human emotional recognition and expression. However, this improvement is marginal, with GPT3.5 demonstrating a significantly higher Mimicry Experiment score than its advanced counterpart. This suggests that between incredibly powerful and advanced models, improvements only create marginally better emotional knowledge and expression.

Bing AI, on the other hand, consistently scored the lowest in all three experiments, accumulating a total score of 427.8/1000, indicating limited capability in mirroring human-like emotional responses. This disparity in performance underscores the potential difference in training datasets, methodologies, and architectural sophistications between Bing AI and more advanced models like Chat GPT4.

In total, Open AI model LLMs (GPT3.5 andChatGPT4) demonstrated a dominant performance over LaMDA (Google Bard), and BLOOM (Hugging Chat). Bing-AI is an Open AI-based model, but its general lack of ability to compete with the other two indicates that it is a far less emotionally sophisticated model.

Individual Emotional Analysis

During the Comprehension Experiment, there was not much variety in the average scores of emotional expression. Joy's average score was 45.68/60, Sadness's was 42.04/60, Fear's was 45.64/60, Anger's was 46.76/60, and Disgust's was 45.56/60. These results indicate that LLMs are worse at interpreting sadness than any other of the key emotions. However, most of the average-lowering score came on behalf of Bing AI's low sadness score, and once this outlier is accounted for, Sadness stands at a similar average to the other emotions.

During the Mimicry Experiment, Joy's average score was 39.12/60, Sadness's was 44.36/60, Fear's was 42.40/60, Anger's was 35.36/60, and Disgust's was 35.24/60. This is fascinating for a few reasons. First, the overall average emotional score was significantly lower for the Mimicry Experiment, indicating one of two possibilities. Either judges are less adept at filtering through emotional impacts of stories or LLMs currently face greater challenges in the mimicry of emotional stimuli. Second, while Sadness received the lowest score during the Comprehension Experiment, it ranks as an uncontested first in the Mimicry Experiment. By contrast, Anger, which had the highest score on the Comprehension Experiment, finished tied with Disgust for last on the Mimicry Experiment. This suggests the emotional

capability associated with mimicry uses a different assessment methodology than interpretation, as indicated by the inverted rankings. Exploring this discrepancy in greater detail should be the subject of future experimentation.

Consistency and Variability in Scoring

The Comprehension Experiment demonstrated a relatively uniform scoring, with rankings within one standard deviation of the mean and no drastic variability from emotion to emotion for each LLM. This suggests a general consistency among participants in evaluating the LLMs' emotional acuity, providing a reliable foundation for interpreting the models' capabilities.

In contrast, the Mimicry Experiment exhibited substantial inconsistencies, both in individual rankings and in emotional rankings per AI. The range in individual rankings was broad, highlighted by significant outliers, revealing a divergence in perception and expectation among participants regarding emotional representation by the LLMs. For instance, the inconsistent scores in emotions, such as Hugging Chat scoring above average in Sadness but record low in Disgust, emphasize the inherent challenge in uniformly quantifying emotional expressions across different emotional categories. Repetition of the Mimicry Experiment may be necessary for more concrete conclusions on the emotional mimicry skills of an AI.

Another possibility explaining the variability of Mimicry Experiment is a lack of consistency on the part of the LLM. Because mimicry is a significantly more complicated emotional process than description, a struggling LLM could be the reason for such discrepancies. This is supported by the distinct types of stories each LLM generated. Google Bard did not seem to have a template for creating these stories, as they were instead highly inconsistent and featured wide ranges of different characters and settings. Hugging Chat's low disgust score may have been attributable to a similar dynamic. The macabre nature of the disgust story was designed to elicit a visceral reaction from a human reader. However, Hugging Chat saw this story as too graphic and against its guidelines to reproduce, refusing to participate in the experiment. It took many additional prompts to finesse the chatbot into actually giving a summary of the story, and as evidenced by the experiment, it was significantly worse.

However, this assumption is disproven by the relative consistency of the Generation Experiment. One would expect generation to be the most complex and thus the most inconsistent of the three emotional metrics. Yet results for the Generation Experiment were remarkably consistent, with script rankings almost proportionately representing final rankings. The underlying reason behind discrepancies in interpretation, generation, or both of LLM emotional mimicry is a topic that should be explored further in future literature.

Implications of Emotional Variability

The findings from Mimicry Experiment, illustrating the variability in emotional scoring, underline the intrinsic subjective nature of emotions and the consequential challenges in modeling them. The disparities in scoring among participants might be attributed to their individual emotional perceptions, understandings, and experiences, emphasizing the importance of diversifying training datasets to encompass a broad spectrum of emotional expressions and interpretations.

Experimentation Difficulties

LLMs frequently failed to cooperate with word limit and content restrictions, meaning that additional prompting was necessary to obtain a valid response. This may have skewed AI responses, as indicated by the above example of Hugging Chat and Mimicry Experiment Disgust. Due to the anonymous nature of the survey, it was impossible to reach out and corroborate submissions. Two participants submitted seemingly random results greater than 2 standard deviations from the mean on the Mimicry Experiment. These results were rejected and the effective sample size of the Mimicry Experiment reduced to 10. No such difficulties were noted in either the Comprehension or Generation Experiments. P value calculations were attempted for each of the three experiments, but such calculations were unsuccessful due to the low sample size. Future experiments with access to a larger respondent pool may better assess the statistical significance of these results through P-value calculations.

Conclusion

The overarching inference drawn from this study is the evident advancement in the capacity of newer LLMs like ChatGPT4 to mirror human emotions more accurately compared to less emotionally sophisticated models like Bing AI. The variability in scores among individual participants and across different emotional categories underscores the inherent challenges in modeling and quantifying emotions due to their subjective nature.

As expressed in the literature review, the LLMs' capacity to demonstrate emotion was at near-human levels. Judging participants indicated through separate communication that some LLMs were indistinguishable from a human response. This experiment supports the general literary trend that LLMs are capable of portraying emotion on a human level. Fully emotionally mature and possibly independently emotional AI could be a mainstay in the near future.

Recommendations for Future Research

Future studies should consider a more extensive and diverse participant pool to mitigate biases and obtain a more comprehensive understanding of human emotional representation. Additionally, exploring the impact of different training datasets and methodologies on the emotional acuity of

LLMs will provide deeper insights into the development of more emotionally responsive and intelligent AI models. No human writing was used as a baseline for this experiment, so future literature can answer the question of whether modern LLMs are more emotionally adept than the average human writer. Finally, investigating the ethical considerations and implications of developing emotionally intelligent AI is crucial to guide responsible innovation in this domain.

Published: February 21, 2024 EDT.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at http://creativecommons.org/licenses/by/4.0 and legal code at http://creativecommons.org/licenses/by/4.0/legalcode for more information.

References

AlBadani, B., Shi, R., & Dong, J. (2022, January 14). *A novel machine learning approach for sentiment analysis on Twitter incorporating the universal language model fine-tuning and SVM*. MDPI. https://www.mdpi.com/2571-5577/5/1/13

Alloghani, M., Thron, C., & Subair, S. (2022). Cognitive computing, emotional intelligence, and artificial intelligence in Healthcare. *Artificial Intelligence for Data Science in Theory and Practice*, 109–118. https://doi.org/10.1007/978-3-030-92245-0_5

Bridges, K. M. B. (1932). Emotional development in early infancy. *Child Development*, *3*(4), 324. https://doi.org/10.2307/1125359

Casas, J., Torche, S., Daher, K., Mugellini, E., & Khaled, O. A. (2021, January 7). *Emotional paraphrasing using pre-trained language models | IEEE*. IEEE Xplore. https://ieeexplore.ieee.org/document/9666309

Cheshin, A., Rafaeli, A., & Bos, N. (2011). Anger and happiness in virtual teams: Emotional influences of text and behavior on others' affect in the absence of non-verbal cues. *Organizational Behavior and Human Decision Processes*, *116*(1), 2–16. https://doi.org/10.1016/j.obhdp.2011.06.002

Churchill, K., & Lipman, L. (2016). Early childhood social and emotional development: Advancing the field of measurement. *Journal of Applied Developmental Psychology*. https://www.sciencedirect.com/science/article/pii/S0193397316300053?via%3Dihub

Ekman, P. (1991). *Universal Emotions*. Psycology.

Frey, C. B., Eckel, C. C., Kabir, G., Steinhubl, S. R., Hinton, G., Darcy, A. M., Obermeyer, Z., Topol, E., Verghese, A., Naylor, C. D., & Cohen, I. G. (2019, November 18). *Artificial Intelligence and the future of psychiatry: Insights from a global physician survey*. Artificial Intelligence in Medicine. https://www.sciencedirect.com/science/article/abs/pii/S0933365719306505

Goswamy, T., Singh, I., Barkati, A., & Modi, A. (2020, December). *Adapting a language model for controlled affective text generation*. ACL Anthology. https://aclanthology.org/2020.coling-main.251/

Grose, J. (2011). Genuine versus deceptive emotional displays. *EPSA Philosophy of Science: Amsterdam 2009*, 77–88. https://doi.org/10.1007/978-94-007-2404-4_8

Heaton, C. T., & Schwartz, D. M. (2020, October 1). Language models as emotional classifiers for textual conversation. *Proceedings of the 28th ACM International Conference on Multimedia*. https://doi.org/10.1145/3394171.3413755

Ho, M.-T. (2022). What is a Turing test for emotional AI? *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01571-3

Holodynski, M. (2008). Milestones and mechanisms of emotional development. *Emotions as Bio-Cultural Processes*, 139–163. https://doi.org/10.1007/978-0-387-09546-2_7

Jack, R. E., Garrod, O. G. B., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, *24*(2), 187–192. https://doi.org/10.1016/j.cub.2013.11.064

Jo, E., Epstein, D. A., Jung, H., & Kim, Y.-H. (2023, April 19). Understanding the benefits and challenges of deploying conversational AI leveraging large language models for Public Health Intervention. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581503

Kollareth, D., Shirai, M., Helmy, M., & Russell, J. A. (2022). Deconstructing disgust as the emotion of violations of body and soul. *Emotion*, *22*(8), 1919–1928. https://doi.org/10.1037/emo0000886

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023, October 20). *Large language models understand and can be enhanced by emotional stimuli*. arXiv.Org. https://arxiv.org/abs/2307.11760

Merriam, G. (2022). If A.I. Only Had a Heart: Why Artificial Intelligence Research Needs to Take Emotions More Seriously. *Journal of Artificial Intelligence and Consciousness*, *9*(1), 73–91. https://doi.org/10.1142/s2705078521500120

Moshkin, V., Konstantinov, A., & Yarushkina, N. (2022). Application of the bert language model for sentiment analysis of social network posts. *Artificial Intelligence*, 274–283. https://doi.org/10.1007/978-3-030-59535-7_20

Nasiri, Y. (2023, April 18). *A language-model-based chatbot that considers the user's personality profile and emotions to support caregivers of people with dementia*. BYU ScholarsArchive. https://scholarsarchive.byu.edu/etd/9862/

OpenAI. (2023). *ChatGPT4 technical report*. arXiv.org. https://arxiv.org/abs/2303.08774

Patel, S. C., & Fan, J. (2023, January 1). *Identification and description of emotions by current large language models*. bioRxiv. https://www.biorxiv.org/content/10.1101/2023.07.17.549421v1.abstract

Prochazkova, E., & Kret, M. E. (2017). Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion. *Neuroscience & Biobehavioral Reviews*, *80*, 99–114. https://doi.org/10.1016/j.neubiorev.2017.05.013

Rosenfeld, A. (2021, May). *Better Metrics for evaluating explainable artificial intelligence*. ResearchGate. https://www.researchgate.net/publication/349111351_Better_Metrics_for_Evaluating_Explainable_Artificial_Intelligence_Blue_Sky_Ideas_Track

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 673–705. https://doi.org/10.1007/s10458-019-09408-y

Russell, P. S., & Giner-Sorolla, R. (2011). Social justifications for moral emotions: When reasons for disgust are less elaborated than for anger. *Emotion*, *11*(3), 637–646. https://doi.org/10.1037/a0022600

Russell, P. S., & Giner-Sorolla, R. (2013). Bodily moral disgust: What it is, how it is different from anger, and why it is an unreasoned emotion. *Psychological Bulletin*, *139*(2), 328–351. https://doi.org/10.1037/a0029319

Santhanam, S., & Shaikh, S. (2019, November 25). *Emotional neural language generation grounded in situational contexts*. arXiv.Org. https://arxiv.org/abs/1911.11161

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023, September 21). *Personality traits in large language models*. arXiv.Org. https://arxiv.org/abs/2307.00184

Simmons, G. (2023, June 17). *Moral mimicry: Large language models produce moral rationalizations tailored to political identity*. arXiv.Org. https://arxiv.org/abs/2209.12106

Wang, X., Li, X., Yin, Z., Wu, Y., & Jia, L. (2023, July 28). *Emotional intelligence of large language models*. arXiv.Org. https://arxiv.org/abs/2307.09042

Zheng, C., Sabour, S., Wen, J., Zhang, Z., & Huang, M. (2022, February). *AugESC: Dialogue augmentation with large language models for emotional support conversation*. NASA/ADS. https://ui.adsabs.harvard.edu/abs/2022arXiv220213047Z/abstract