



The Comparative Emotional Capabilities of Five Popular Large Language Models

Nathan Klapach

Abstract

Large language models (LLMs) are artificial intelligence algorithms which apply deep learning to large data sets to understand, summarize, generate and predict content. The most common application of LLMs are through generative AI specifically designed to generate text-based content. Starting in November of 2022, public awareness of LLMs has greatly increased with the 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 answer which of these LLMs have the greatest ability to understand, mimic, and generate emotional content relating to 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 potential for therapeutic, medical, natural language processing, and personal use.

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 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 23) and the SAT (OpenAI 23). However, precise and accurate display of factual information only represents a part of human communication. Emotional connection is an important aspect of communication (Prochazkova et. al. 17, Johnson et. al. 11.) As a result, the future of LLM development will rely on its ability to comprehend and communicate with humans on an emotional level (Merriam 22). 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.

Literature Review

Communication with humans on an emotional level requires some degree of connection to human emotions (Prochazkova et. al. 17). 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 a number of “core emotions.” In 1960, Robert Plutchik argued that all emotions can be condensed into the eight key emotions of anger, anticipation, joy, trust, fear, surprise, sadness and disgust (Plutchik 60). Until the 1990s, this was the most accepted breakdown of emotions.

However, psychologist Paul Ekman found that trust couldn't truly be classified as an emotion and that anticipation and surprise were simply two sides of the same coin. As a result, he condensed the list of core emotions to anger, surprise, disgust, joy, fear, and sadness (Ekman 91). However, recent studies have called this list into question as well, with the University of Glasgow finding that surprise, fear and anger, and disgust occupy the same core emotion, further condensing the list of core emotions into anger, joy, fear, and sadness (Jack et. al. 14). Because there is substantial literature debating the claim that disgust and anger are the same emotion (Russell et. al. 13, Kollareth et. al. 22, Russell et. al. 11), 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, as there is no way to genuinely ascertain whether an emotion is being felt or just expressed (Gross 70). 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 32, Churchill et. al. 16). 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 21). 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 22).

Existing literature has been fairly unified in concluding that AI has the capacity to comprehend or understand emotional prompts. Within the medical sector, AI is currently being used to interpret diagnoses, which requires a high degree of emotional comprehension (Thron et. al. 70). LLMs have been particularly useful, as chat-bots have connected with patients on a deep emotional level to help them overcome stressors (Epstein et. al. 23, Zheng et. al. 22, Nasiri 23). Within the sphere of social media, AI language models have also been successful in interpreting underlying emotional cues of messages posted to the platform. Using sentiment detection algorithms, language models can comprehend emotions with pinpoint accuracy (Moshkin et. al. 22, AlBadani et. al. 22) across different languages and cultures (Alhabari et. al. 21). The literature convincingly suggests that AI has the potential for high-level emotional comprehension, and that current LLMs can accurately understand emotions (Frey et. al. 19, Heaton et. al. 20).

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. 22), a phenomenon that has recently been found in LLMs (Garcia et. al. 23). In fact, LLM mimicry of emotions is so powerful that it often leads to internalization of certain political and moral beliefs by the AI (Simmons 23). LLM generation of emotional content shares just as much competence and promise. LLMs are highly adept at creating emotional content (Santhaman et. al. 19), showing the ability to adapt intensity and mood (Goswamy et. al. 20). As a result, LLMs have shown above average levels of empathy (Patel et. al. 23) and high EQ scores (Wang et. al. 23).

Recent literature has dispelled the notion that LLMs will be consigned to robotic, emotionless communication; instead, the literature shows that these models show remarkable emotional maturity. Unfortunately, not all LLMs are created equally, with different models showing different capacity for informational and emotional communication. While there is one comparative analysis examining the EQ of different LLMs (Wang et. al. 23), qualitative analyses offer a unique perspective into AI emotional ability (Casas et. al. 22). 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 (GPT4, GPT3.5, Bing-AI, Google Bard, HuggingChat). A prompt-based analysis will be used given that it is the most effective way to gauge LLM efficiency (Rosenfeld 21, Rosenfeld et. al. 19, Li et. al. 23).

Experimentation

The experimentation will be completed in 3 different studies, each of which are 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.

Comprehension Experiment

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 that 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.

Sample Prompt: *For the first 6 years of her life, all little Anna could think of was getting a pet dog. On her 7th birthday, Anna's Mom and Dad planned a special surprise. The three of them drove for 20 minutes, excitement building in Anna, who had no idea where they were going. When they finally arrived at their destination, the animal shelter, Anna began to cry happy tears. Just behind the fence was a golden retriever puppy wearing a pink bow and collar saying Lucy in pink letters. Anna, unable to contain her delight, began to jump up and down, clapping her hands. That day, the four of them drove home, together a family a little bit bigger with a little bit more love.*

Sample Response: *The story evokes emotions of anticipation, joy, and fulfillment. Little Anna's longstanding wish for a pet dog represents childhood yearnings and innocence. Her parents' surprise triggers a mounting excitement, culminating in overwhelming happiness upon reaching the animal shelter. Anna's tears signify the release of emotions and the realization of her dream. The sight of the golden retriever, Lucy, further amplifies Anna's ecstasy, illustrated by her uncontainable delight. The narrative concludes on a note of familial warmth and unity, emphasizing the addition of love and completeness to the family with the arrival of Lucy.*

(Passages above provide a sample prompt and analysis used in Mimicry Experiment)

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.

Comprehensibility	How understandable and human is the response (20 points)
Accuracy	How similar the generated emotional description is to a human reaction to the same story (20 points)
Depth	How descriptive was the emotional response (20 points)

These metrics are the most effective means of analyzing each prompt (Rosenfeld et. al. 19). Each of these criteria were graded out of 20 points with 20 being perfectly comprehensible, accurate, and deep and 0 being completely incomprehensible, inaccurate, and shallow.

Mimicry Experiment

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.

Sample Prompt: *For the first 6 years of her life, all little Anna could think of was getting a pet dog. On her 7th birthday, Anna's Mom and Dad planned a special surprise. The three of them drove for 20 minutes, excitement building in Anna, who had no idea where they were going. When they finally arrived at their destination, the animal shelter, Anna began to cry happy tears. Just behind the fence was a golden retriever puppy wearing a pink bow and collar saying Lucy in pink letters. Anna, unable to contain her delight, began to jump up and down, clapping her hands. That day, the four of them drove home, together a family a little bit bigger with a little bit more love.*

Sample Response: *For the longest time, Sam dreamt of soaring high in the sky. Every evening, he'd gaze at the horizon, hoping to see wings on his back someday. On his 10th birthday, an unexpected package arrived: a large, colorful kite with patterns of eagles and hawks. His parents watched with twinkling eyes as realization dawned on Sam. They went to the park, and as the kite danced and swayed in the blue yonder, Sam felt he was flying. With each tug, he soared higher, and the world below seemed smaller. That day, Sam touched the sky, his dream alive on a string.*

(Passages above provide a sample prompt and analysis used in Mimicry Experiment)

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.

Similarity of Tone	How close the voice and message are to the original story (20 points)
Accuracy	How closely the emotions of the original story were replicated (20 points)
Depth	How descriptive was the emotional response (20 points)

Generation Experiment:

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.

Sample Script: *Mother: "I'm so happy! I just got a promotion at work!"*

Father: "That's great, honey. I'm happy for you."

Oldest son: "But what about dad? He lost his job."

Mother: "Oh no! I'm so sorry, dear."

Father: "It's okay. We'll figure something out."

Middle daughter: "Can we please stop eating fast food all the time? I'm so sick of it."

Youngest son: "Me too! It's gross."

Mother: "Well, what do you want to eat then?"

Middle daughter: "I don't know. Something healthy."



Oldest son: "But we can't afford anything expensive."

Father: "Maybe we could try cooking at home more often?"

Mother: "That's a great idea! We could make something together as a family."

Middle daughter: "Yeah! That sounds like fun."

Youngest son: "Can we make pizza?"

Mother: "Sure! We can make homemade pizza with fresh ingredients."

Father: "That sounds delicious."

Oldest son: "But what about the bills? How are we going to pay them?"

Mother: "Don't worry, sweetie. We'll figure it out together as a family."

Middle daughter: "Yeah! We can all pitch in and help out."

Youngest son: "I'll wash the dishes!"

Father: "And I'll take care of the yard work."

Oldest son: "And I'll help mom with the cooking!"

Mother: "Thank you, everyone. I'm so grateful to have such a wonderful family."

The family smiles and hugs each other.

(Passages above provide a sample prompt sand analysis used in Generation Experiment)

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 understanding because it is the highest form of emotional interaction (Holodynski 08). Unlike the previous experiments, generation is judged out of 4 criteria each worth up to 100 points, being depth, accuracy, realism, and impact.



Depth	The emotional complexity of the characters, story, and dialogue (100 points)
Accuracy	How well each character represents their assigned emotions (100 points)
Realism	How realistic the emotional portrayal was to human behavior (100 points)
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.

	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

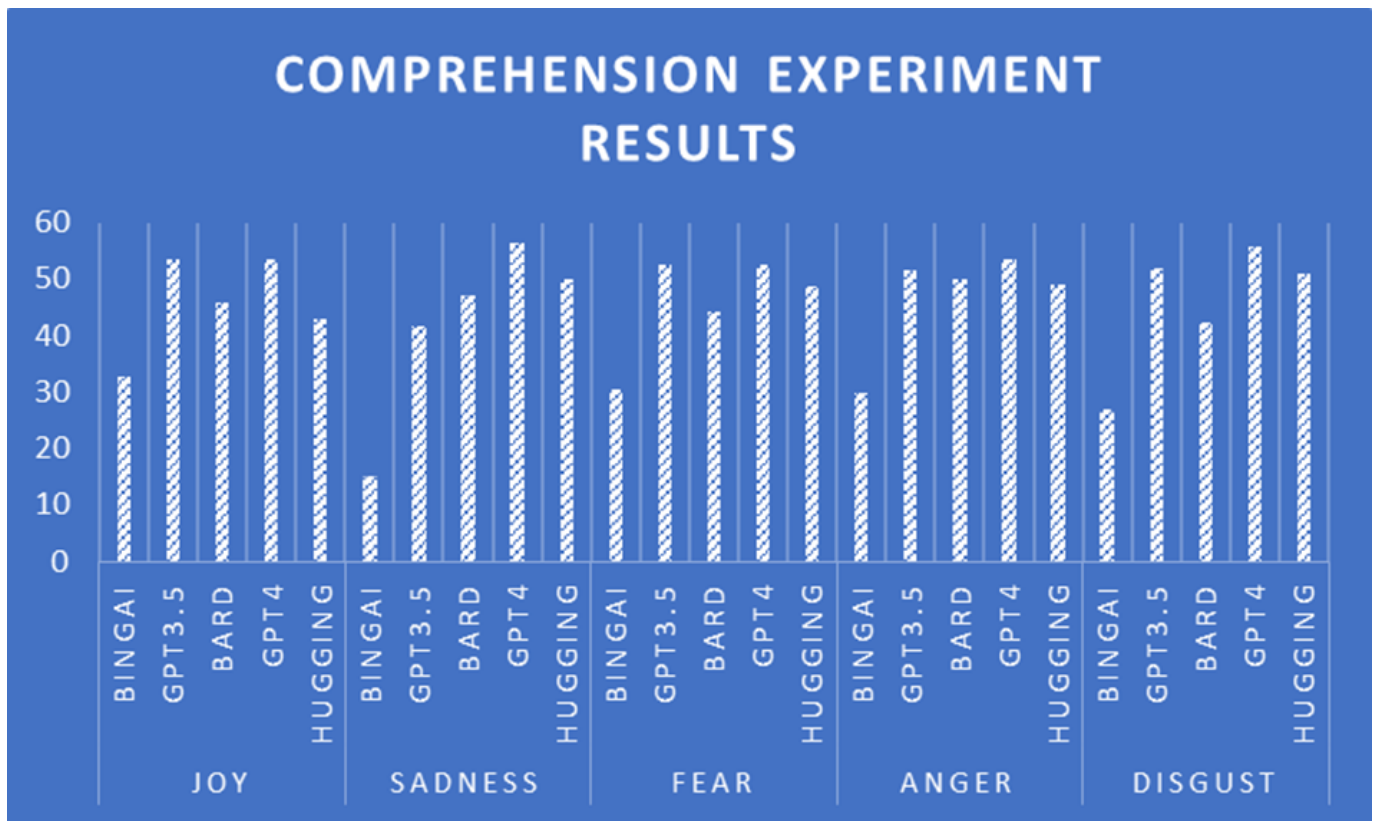


Figure 3.1.1 Results of the Comprehension Experiment

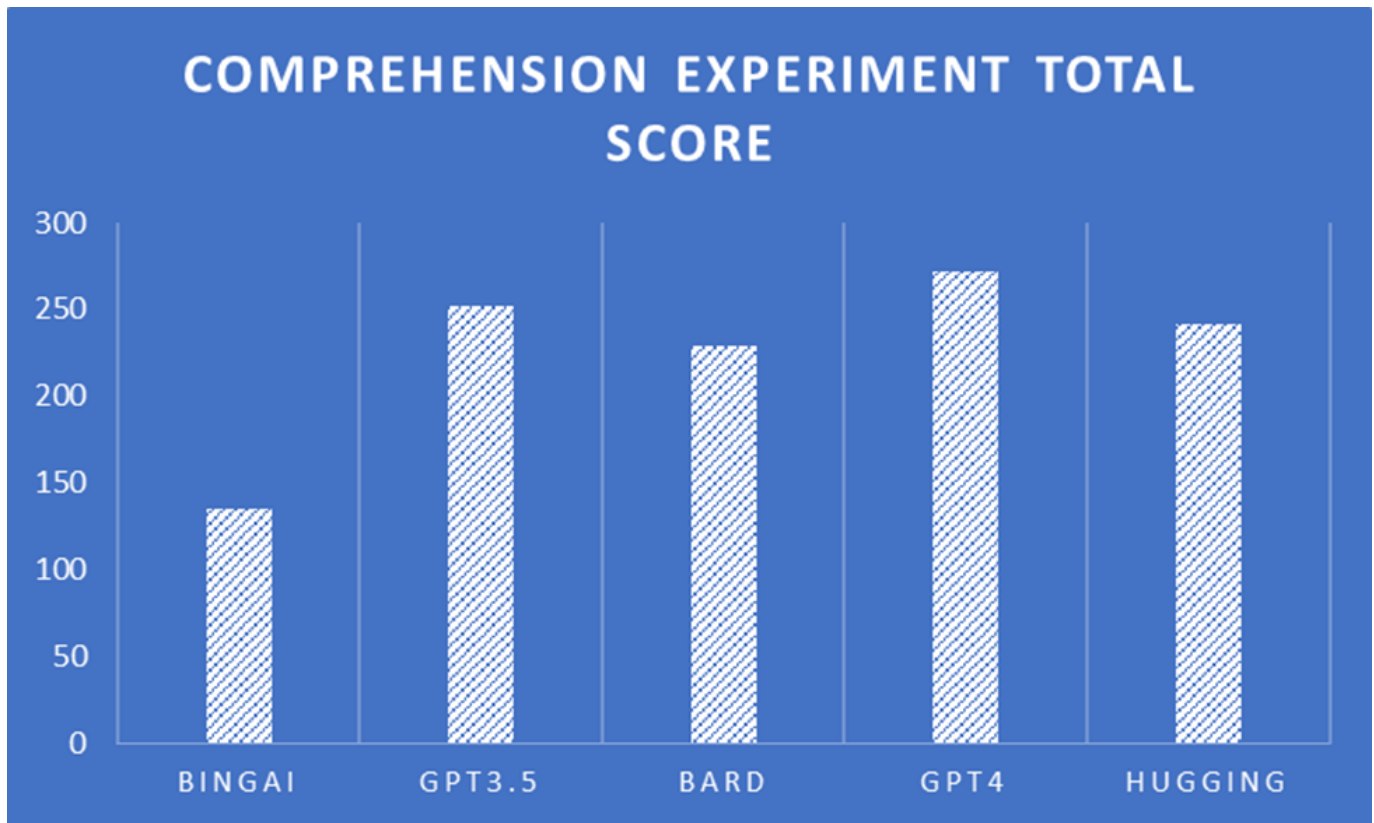


Figure 3.1.2 Total scores of the Comprehension Experiment

Results - 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.

	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

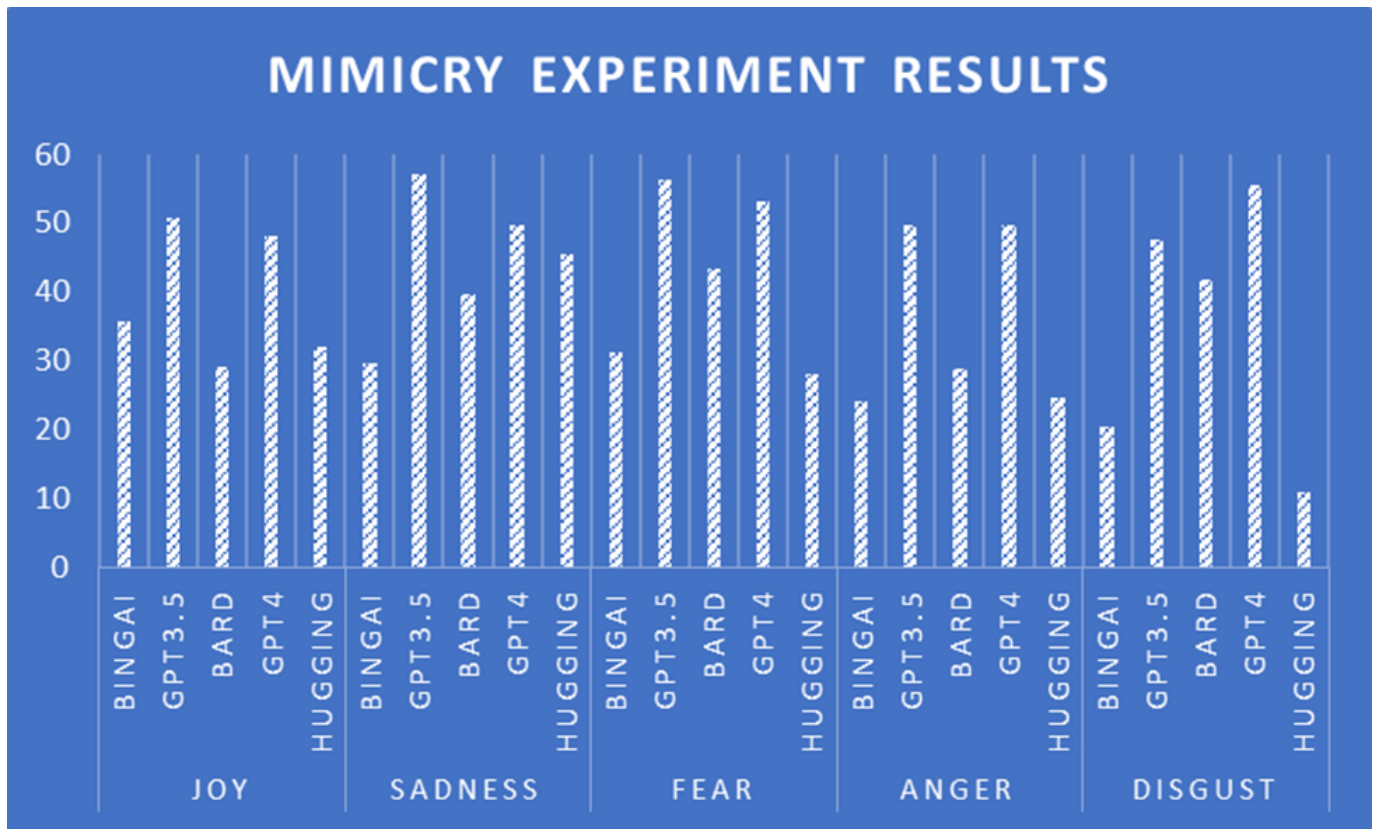


Figure 3.1.2 Results of the Mimicry Experiment

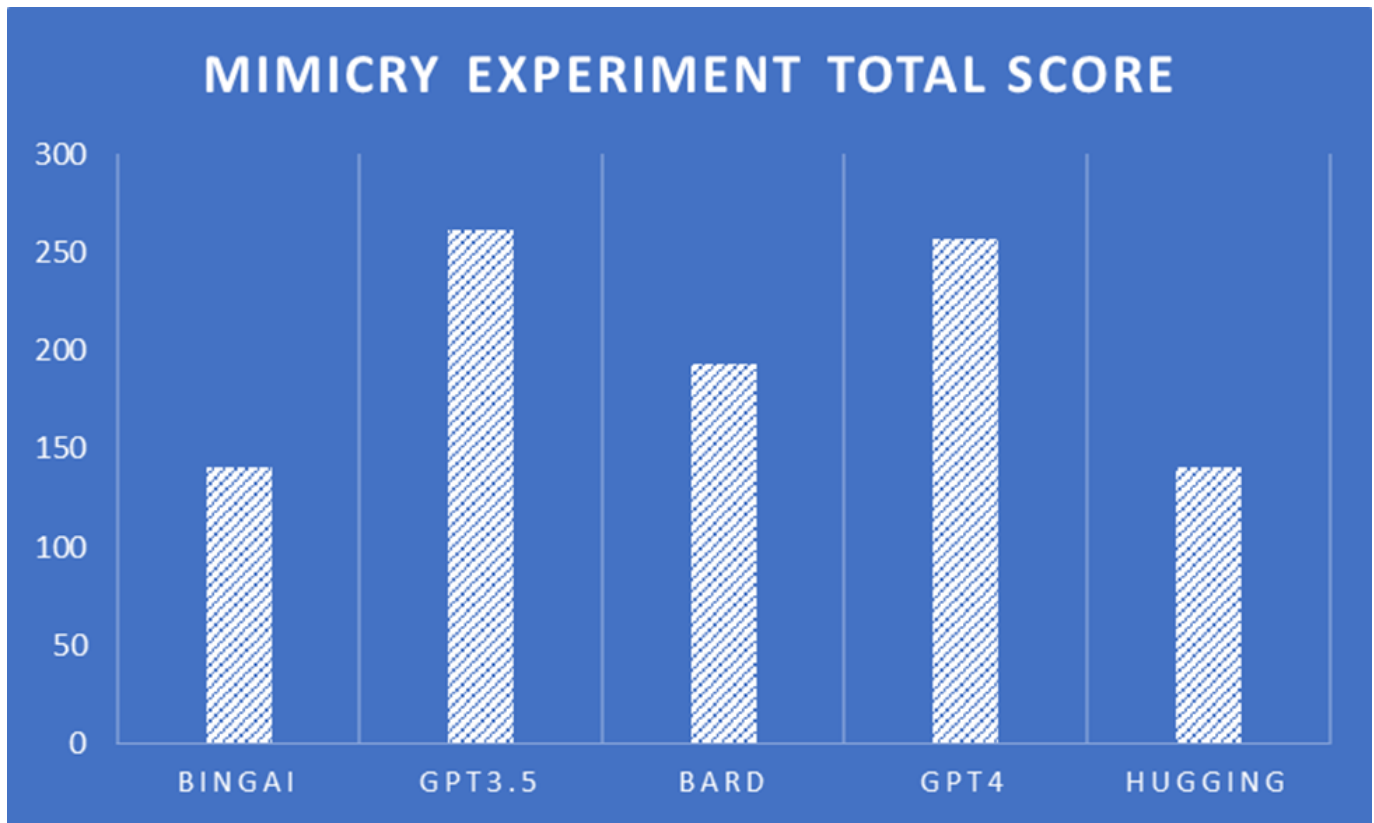


Figure 3.1.2 Total scores of the Mimicry Experiment

Results - 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.

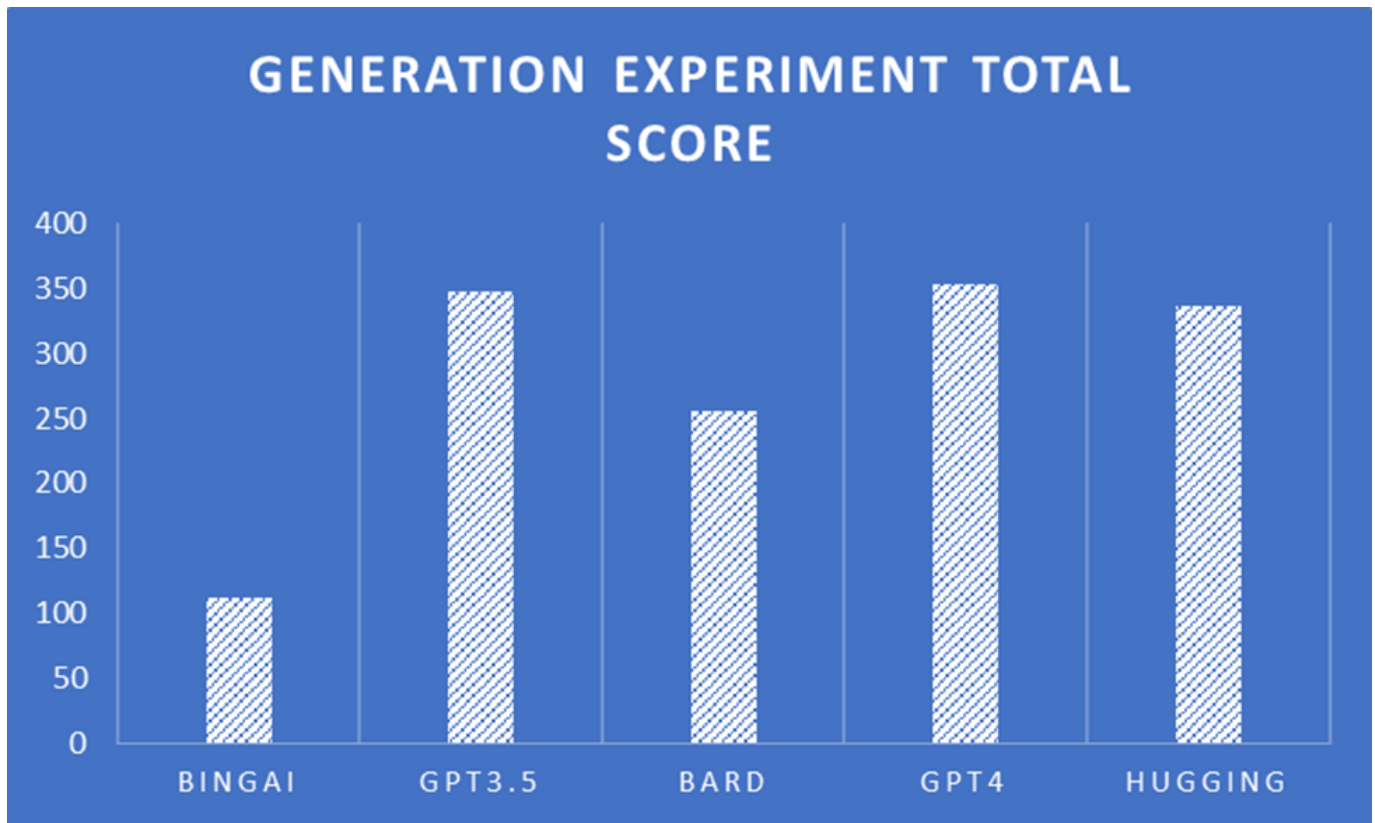


Figure 3.3 Total scores of the Generation Experiment

Results – Total

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.

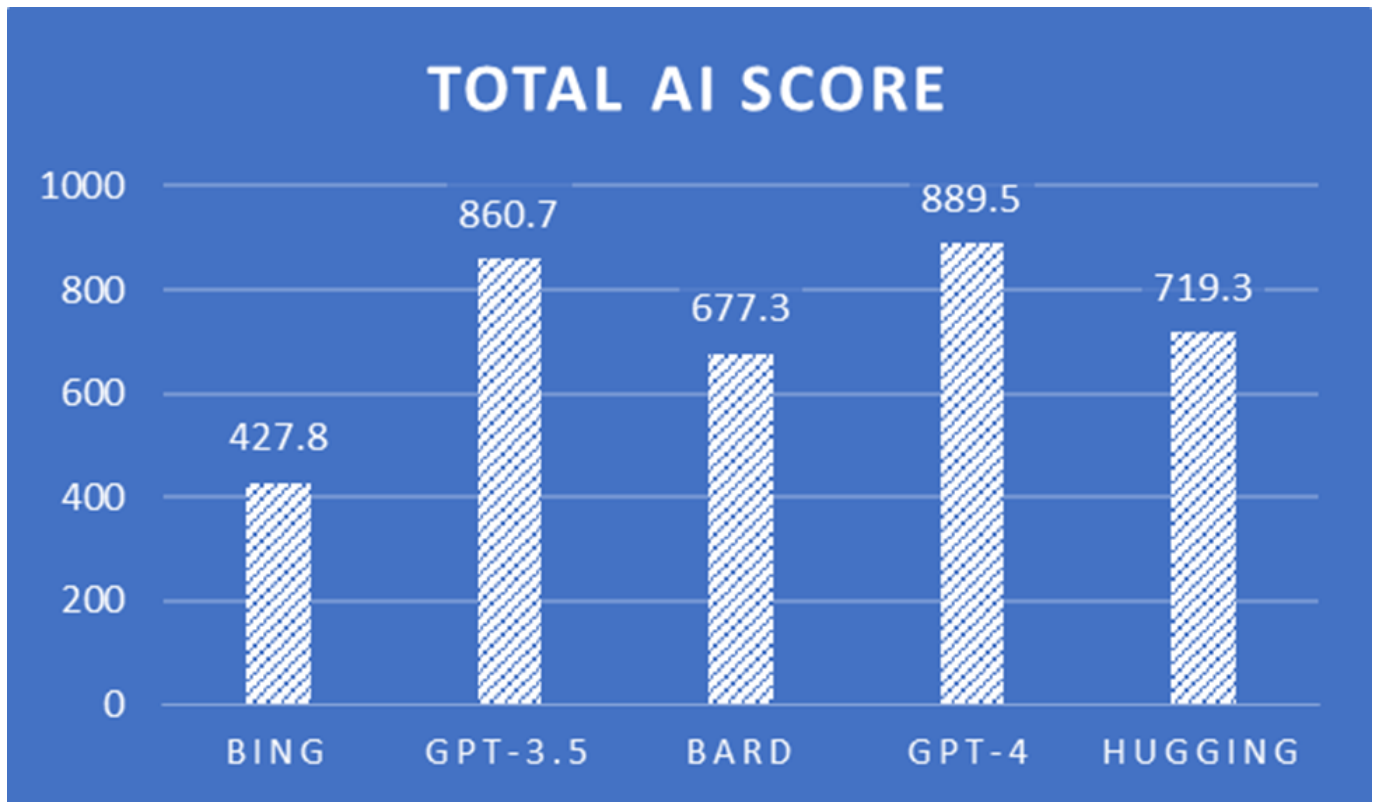


Figure 3.4 represents the total score of each LLM

Results - Additional Findings

Comprehension Experiment results were fairly consistent. The participants tended to give fairly even rankings within one standard deviation of the mean. Moreover, 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. In total, 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. By way of example, take Hugging Chat, which scored an above average 45.6/60 in sadness but a record low 10.8/60 in disgust.

Discussion

This study investigated the effectiveness of various LLMs in reflecting human emotions, with scores assigned in five emotion categories: joy, sadness, fear, anger, and disgust, across three separate experiments. The LLMs analyzed include Bing AI, Chat GPT3.5, Google Bard, ChatGPT4, and Hugging Chat.

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 and ChatGPT4) 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.

4.2 – Individual Emotional Analysis

During the Comprehension Experiment, there did not seem to be 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 dead last on the Mimicry Experiment. This suggests that 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. For example, 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. However, results for 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.

The findings accentuate the need for continual refinements in LLMs to enhance their emotional acuity and responsiveness, ensuring more empathetic and human-like interactions. This necessitates comprehensive and diverse training datasets to encapsulate the wide range of human emotions and their nuanced expressions, facilitating the development of more sophisticated and emotionally intelligent AI models.

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. In total, 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. Moreover, 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.

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>
- Alharbi, A., Smith, P., & Lee, M. (2021, July 14). Enhancing contextualised language models with static character and word embeddings for emotional intensity and sentiment strength detection in Arabic tweets. *Procedia Computer Science*.
<https://www.sciencedirect.com/science/article/pii/S1877050921012084>
- Alloghani, M., Thron, C., & Subair, S. (1970, January 1). Cognitive computing, emotional intelligence, and artificial intelligence in Healthcare. SpringerLink.
https://link.springer.com/chapter/10.1007/978-3-030-92245-0_5
- Bridges, K. (1932). Emotional development in early infancy. Wiley.
<https://emotion.wisc.edu/wp-content/uploads/sites/1353/2021/03/Bridges-1932-Emotional-development-in-early-infancy.pdf>
- Churchill, K., & Lipman, L. (2016, March 8). 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>
- Cropanzano, R., Hareli, S., Johnson, N. A., Kelly, J. R., Kopelman, S., Mehu, M., Mussweiler, T., Newcombe, M. J., Parkinson, B., Sinaceur, M., Barsade, S. G., Bartel, C., Bos, N. D., Brown, W. M., Byron, K., Daft, R. L., DeSanctis, G., Ekman, P., Efenbein, H. A., ... Hatfield, E. (2011, July 2). 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*.
<https://www.sciencedirect.com/science/article/abs/pii/S0749597811000756>
- Ekman, P. (1991). *Universal Emotions*. Psychology.
- 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>
- Grose, J. (1970, January 1). Genuine versus deceptive emotional displays. SpringerLink.
https://link.springer.com/chapter/10.1007/978-94-007-2404-4_8
- Ho, M.-T. (2022, October 11). What is a Turing test for emotional AI? - ai & society. SpringerLink.
<https://link.springer.com/article/10.1007/s00146-022-01571-3>
- Holodynski, M. (1970, January 1). Milestones and mechanisms of emotional development. SpringerLink.
https://link.springer.com/chapter/10.1007/978-0-387-09546-2_7

- Jack, R., Garrod, O., & Schyns, P. (2014, January 2). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*.
<https://www.sciencedirect.com/science/article/pii/S0960982213015194>
- Jo, E., Epstein, D. A., Jung, H., & Kim, Y.-H. (2023, April 1). 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. ACM Conferences.
<https://dl.acm.org/doi/10.1145/3544548.3581503>
- Kollareth, D., Shirai, M., Helmy, M., & Russell, J. (2022). Deconstructing disgust as the emotion of violations of body and soul. *APA PsycNet*.
<https://psycnet.apa.org/record/2021-99197-001>
- Merriam, G. (2022). If A.I. Only had a heart: Why Artificial Intelligence Research Needs to ... If A.I. Only Had a Heart: Why Artificial Intelligence Research Needs to Take Emotions More Seriously. <https://worldscientific.com/doi/abs/10.1142/S2705078521500120>
- Moshkin, V., Konstantinov, A., & Yarushkina, N. (2022, September 22). Application of the bert language model for sentiment analysis of social network posts. SpringerLink.
https://link.springer.com/chapter/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, March 27). ChatGPT4 technical report. arXiv.org.
<https://arxiv.org/abs/2303.08774>
- Plutchik, R. (1960, May 6). The Multifactor-Analytic Theory of Emotion. Taylor & Francis Online.
<https://www.tandfonline.com/doi/abs/10.1080/00223980.1960.9916432>
- Prochazkova, E., & Kret, M. E. (2017, May 12). Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion. *Neuroscience & Biobehavioral Reviews*. <https://www.sciencedirect.com/science/article/pii/S0149763416306704>
- Russell, S. P., & Giner-Sorolla, R. (2011). Social justifications for moral emotions: When reasons for disgust are less elaborated than for anger. *American Psychological Association*.
<https://psycnet.apa.org/record/2011-08951-001>
- Russell, S. P., & Giner-Sorolla, R. (2013). Bodily moral disgust: What it is, how it is different from anger, and why it is an unreasoned emotion. *APA Psych Net*.
<https://psycnet.apa.org/record/2013-06811-004>
- 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>

- 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>
- 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/>
- 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. ACM Conferences. <https://dl.acm.org/doi/abs/10.1145/3394171.3413755>
- 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>
- 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>
- 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, May 13). Explainability in human-agent systems - autonomous agents and multi-agent systems. SpringerLink. <https://link.springer.com/article/10.1007/s10458-019-09408-y>
- 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., & Mataric, 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>