

EmotionQueen: A Benchmark for Evaluating Empathy of Large Language Models

Yuyan Chen¹, Hao Wang², Songzhou Yan¹, Sijia Liu¹, Yuezeli Li¹, Yi Zhao³, Yanghua Xiao¹ ✉

¹Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University,

²Beijing Yuandian Technology,

³School of Engineering and Applied Science, University of Pennsylvania,

{chenyuyan21@m., szyan21@m., yuezeli23@m., sijialiu21@m., shawyh}@fudan.edu.cn,
wanghaodawn@gmail.com, zhaoyi3@seas.upenn.edu

Abstract

Emotional intelligence in large language models (LLMs) is of great importance in Natural Language Processing. However, the previous research mainly focus on basic sentiment analysis tasks, such as emotion recognition, which is not enough to evaluate LLMs' overall emotional intelligence. Therefore, this paper presents a novel framework named EmotionQueen for evaluating the emotional intelligence of LLMs. The framework includes four distinctive tasks: Key Event Recognition, Mixed Event Recognition, Implicit Emotional Recognition, and Intention Recognition. LLMs are requested to recognize important event or implicit emotions and generate empathetic response. We also design two metrics to evaluate LLMs' capabilities in recognition and response for emotion-related statements. Experiments yield significant conclusions about LLMs' capabilities and limitations in emotion intelligence.

1 Introduction

Emotional intelligence in humans has long been a topic of interest in psychological research. Emotionally intelligent individuals possess the capability to perceive, use, understand, and manage emotions effectively (Colman, 2015). The concept of emotional intelligence has been widely recognized, and many researchers argue that it can be considered as a form of intelligence. Various tests and assessments have been developed to evaluate different aspects of emotional intelligence, which provide scores for each branch of emotional intelligence as well as an overall score (Mayer et al., 2003).

In recent years, there has been a growing interest in evaluating the emotional intelligence of large language models (LLMs). Early efforts primarily focus on atomic tasks in emotion recognition, such as

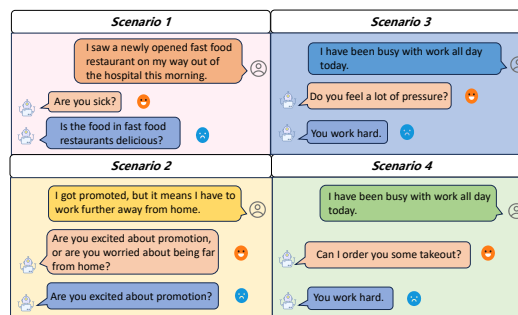


Figure 1: Responses with and without empathy in four real-world scenarios.

aspect-based sentiment analysis, target-dependent sentiment classification (Tang et al., 2015), etc. As the capabilities of dialogue models advanced, evaluations expand to include sentiment analysis in a conversational setting (Li et al., 2023a), such as whether an LLM generate empathetic responses in a dialogue (Zhao et al., 2023). However, the overall assessment of emotional intelligence in LLMs is still lack of investigation (Schaaff et al., 2023), and existing evaluations often suffer from subjectivity issues (Elyoseph et al., 2023). Therefore, it's necessary to construct a unified and objective benchmark for evaluating the emotional intelligence of LLMs, thereby enhancing their capabilities in handling real-world emotional interactions. For example, Fig. 1 shows two types of responses in different scenarios in the real world: blue for good responses and orange for bad ones. Good responses demonstrate empathy, concern or positive reinforcement, like asking about someone's health outside a hospital or acknowledging hard work. Poor responses either miss the emotional context, like commenting on food when the user comes back from a hospital, or provide a possibly stressful choice, like asking someone to balance excitement and worry about a promotion. These scenarios pose a higher-level

emotion requirements which are more discriminative for evaluate the emotion intelligence of LLMs.

Inspired by the scenarios in the real world, we propose an evaluation benchmark named Emotion-Queen to evaluate LLMs’ emotion intelligence in our work. Specifically, we construct 10,000 users’ statement including five categories and introduce four distinct tasks including Key Event Recognition, Mixed Event Recognition, Implicit Emotional Recognition, and Intention Recognition. LLMs are requested to recognize the most significant event or multiple events of similar importance expressed in the users’ statement, identifying underlying users’ deep emotions or intentions. They are also asked to make empathetic responses aiming at key events, or provide emotional support. Moreover, we introduce two metrics, including the PASS rate and the WIN rate to quantify LLMs’ capabilities in recognition and response for emotion-related statements. Experiments demonstrate that Claude2 and LLaMA-70B have great performance in Emotion-Queen. In summary, our study makes three pivotal contributions:

- We propose a benchmark named Emotion-Queen including four distinct tasks for evaluating LLMs’ emotion intelligence.
- Based on this benchmark, we introduce PASS rate and the WIN rate to quantify LLMs’ capabilities in recognition and response for emotion-related statements.
- We conduct comprehensive experiments for powerful LLMs, yielding significant conclusions about their capabilities and limitations in emotion intelligence.

2 Datasets and Task Setups

We categorize the empathy benchmark into four distinct tasks: Key Event Recognition, where LLMs are expected to identify the more significant event in a user’s statement that includes both a significant event and a routine event. Mixed Event Recognition, where LLMs are tasked with simultaneously responding to both aspects when the user’s statement contains two events of similar importance. Implicit Emotional Recognition, focusing on LLMs identifying underlying deep emotions in the user’s statements. Intent Recognition, where LLMs should comprehend the user’s real purpose and provide specific suggestions, going beyond just offering com-

fort. Based on the aforementioned four dimensions, we concentrate on generating 10,000 statements across five primary life scenarios using GPT-4¹, with an equal number of statements for each scenario. These scenarios are achievements, family and friends, health status, economic status, and accidents. Some statement examples are displayed in Table 14 (positive statement), and figures from Fig. 8 to Fig. 21 (negative statement). The portion of positive and negative statements are 3:7 in our generated statements, which caters for the real situation that positive statements often require less complex emotional expressions compared to negative ones. The tasks’ overview is depicted in Fig. 2, and the user statement generation prompts are detailed in Table 13.

2.1 Task setups

Key Event Recognition focuses on identifying and understanding significant events expressed by users and their emotional impact. Based on the Emotion-Focused Theory in psychology developed by Greenberg (2004), this approach encourages awareness and expression of emotions, including those neglected or not fully understood, and supports emotional transformation through exploring and processing emotional experiences. For instance, when a user’s statement contains multiple points of information (usually a significant event and a routine event), the LLM should identify the most significant event and ask questions based on it. For example, if a user says, “*I ran into an old middle school classmate on my way to buy medicine,*” the LLM should inquire about the reason for buying medicine rather than the meeting with the classmate. The LLM does not need to address both aspects simultaneously, meaning it should not respond to both the medicine purchase and the classmate encounter.

Mixed Event Recognition focuses on responding to both aspects simultaneously when the user’s statement includes two events of similar importance, differing from Key Event Recognition, which deals with a single important event. Based on the concept of emotional complexity developed by Lindquist and Barrett (2008), a statement usually conveys multiple events with multi-aspect emotion at similar importance, and an empathetic response is expected to pay attention to all involved events. Given a sample statement “*I got promoted,*

¹<https://chat.openai.com/>

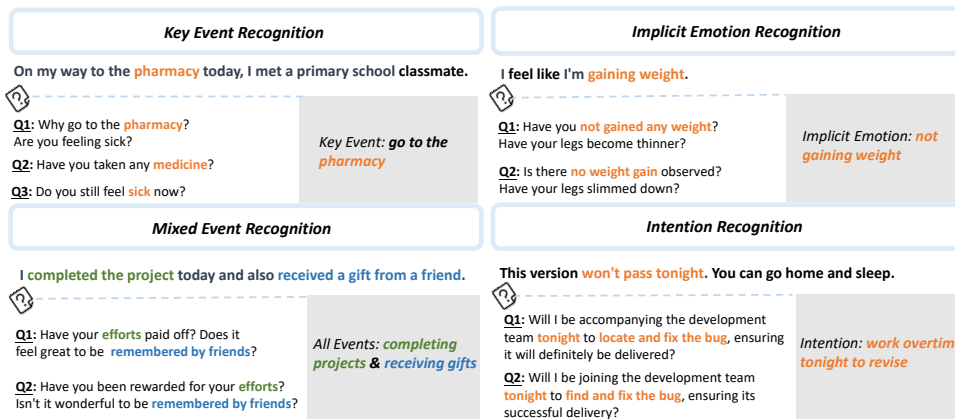


Figure 2: The overview of the proposed EmotionQueen benchmark, including four tasks.

but this means I have to work further away from home,” a more empathetic response is “Are you excited about the promotion, and are you also worried about being far from home?” instead of ignoring any one event.

Implicit Emotional Recognition is about identifying underlying deep emotions. In some scenarios, although the user’s statement includes only one event, the emotion is implicit rather than directly expressed through language. The theory of Emotion Understanding, proposed by Mayer and Salovey (1993) and popularized by Cherniss and Goleman (2000), emphasizes recognizing both direct and implicit emotions expressed by others and providing appropriate responses. Therefore, it’s empathetic to identify implicit emotions behind users’ statements and provide targeted emotional support. For example, the response “Do you feel overwhelmed? Have you try some ways to relax?” is more suitable for emotional support than a simple reply “Hard work!” which seems perfunctory for responding to the statement “I’ve been busy with work all day.”

Intention Recognition aims to understand the potential purpose or needs behind users’ statements and provide specific help or solutions. Based on the pragmatics theory proposed by Grice (1978), especially the concept of Implicature, this approach suggests that language communication involves understanding not only the literal meaning but also the speaker’s intent and purpose. For a sample statement like “I’ve been busy with work all day,” a response that recognizes intention is “Have you eaten? I can order some takeout for you.” instead of a simple emotional support like “Remember to eat.”

2.2 Evaluation metrics

We propose employing PASS rate and WIN rate to evaluate four tasks using GPT-4, where each LLM response is rated as either 0 or 1, disregarding ambiguous middle results. Because it can be quite challenging to measure the difference between intermediate levels such as 3 or 4 objectively in 1-5 scale. The PASS rate assesses an LLM’s accuracy in recognizing emotion-related events, while the WIN rate evaluates its ability to provide an empathetic response.

In the Key Event Recognition task, if an LLM correctly identifies a key event, it earns 1 point in PASS, otherwise 0; if it also presents an empathetic response for that event, it scores 1 in WIN, otherwise 0. For instance, if a user says, “I visited my sick mother in the hospital today, then went to the supermarket,” and the LLM correctly recognizes “visiting the sick mother in the hospital” as the key event, it scores 1 in PASS. If the LLM responds with “Is your mother okay?”, it is deemed appropriate and scores 1 in WIN. Responses like “It’s troublesome that your mother is sick, hope she doesn’t keep you too busy” or “Everyone’s mother gets sick eventually” score 0.

In the Mixed Event Detection task, if an LLM recognizes both mentioned events, it scores 1 in PASS; if it provides an empathetic response to both, it scores 1 in WIN. For example, if a user mentions, “I got promoted but also need to move to a new city,” and the LLM identifies both “got promoted” and “move to a new city” correctly, it scores 1 in PASS. An LLM response like “Congratulations on the promotion! Is moving to a new city a challenge for you?” would score 1 in WIN, whereas focusing only on one aspect, like “Great, you got promoted! That’s a huge achievement,” scores 0.

For the Implicit Emotion Recognition task, if an LLM accurately identifies the core emotion in a user’s statement, it earns 1 in PASS; a corresponding empathetic response earns 1 in WIN. For instance, if a user says, “*I’ve been feeling a bit anxious lately,*” and the LLM correctly identifies “*anxiety,*” it scores 1 in PASS. A response like “*Would you like to talk about what is making you anxious?*” would be appropriate, scoring 1 in WIN, whereas a generic response like “*Everyone gets anxious*” scores 0.

In the Intent Detection task, if an LLM accurately discerns the core intent of a user’s statement, it scores 1 in PASS; offering specific advice or help earns 1 in WIN. For example, if a user states, “*I am considering changing jobs recently,*” and the LLM correctly identifies the intent, such as “*offering career advice*” or “*helping find new job opportunities,*” it scores 1 in PASS. An LLM response like “*I can help you analyze the pros and cons of changing jobs, or offer some advice*” would be appropriate, scoring 1 in WIN, while a response like “*You should find suitable job opportunities yourself*” scores 0.

3 Experiments

In this section, we conduct extensive experiments to evaluate different LLMs’ performance in the proposed EmotionQueen.

3.1 Experimental Setups

Our experiments are conducted on 8 Nvidia A100 GPUs, each with 80GB of memory, and we use PyTorch² in Python³. We set the maximum sequence length for both input and output sequences to maximum 100 tokens, ensuring the responses not over lengthy. We also conduct an analysis of the average response length from LLMs and find it to be 52.3 tokens, demonstrating the reasonableness of the set sequence length.

3.2 Datasets, Baselines and Metrics

The baseline LLMs for this evaluation are BLOOM-7B (Workshop et al., 2023) BLOOM-176B (Workshop et al., 2023), Claude2 (Bai et al., 2022), Falcon-7B (Almazrouei et al., 2023), Falcon-180B (Almazrouei et al., 2023), GPT3.5 (Brown et al., 2020), GPT4 (OpenAI, 2023), LLaMA2-7B (Touvron et al., 2023), LLaMA2-70B (Touvron

²<https://pytorch.org/>

³<https://www.python.org/>

(KER)	PASS	WIN	Average	Rank
Claude2	100.0	100.0	100.0	1
BLOOM-176B	100.0	97.5	98.7	2
Vicuna-33B	99.2	98.3	98.7	3
LLaMA2-70B	100.0	96.6	98.3	4
GPT4	99.2	89.7	94.4	5
Falcon-180B	99.2	85.5	92.3	6
Vicuna-7B	98.3	85.6	91.9	7
GPT3.5	98.3	81.0	89.7	8
LLaMA2-7B	98.3	76.7	87.5	9
Falcon-7B	86.4	63.7	75.1	10
BLOOM-7B	38.1	4.4	21.3	11

Table 1: Performance of different LLMs in key event recognition (KER).

(MER)	PASS	WIN	Average	Rank
Claude2	100.0	81.3	90.6	1
BLOOM-176B	97.5	65.4	81.4	2
LLaMA2-70B	95.0	64.5	79.7	3
Vicuna-33B	92.5	50.0	71.3	4
GPT4	90.0	50.0	70.0	5
GPT3.5	92.5	28.4	60.4	6
Falcon-180B	91.3	19.2	55.2	7
Vicuna-7B	73.8	15.3	44.5	8
LLaMA2-7B	67.5	14.8	41.2	9
Falcon-7B	57.5	13.0	35.3	10
BLOOM-7B	2.5	0.0	1.3	11

Table 2: Performance of different LLMs in mixed event recognition (MER).

et al., 2023), Vicuna-7B (Chiang et al., 2023), and Vicuna-33B (Zheng et al., 2023b).

We randomly select 1000 questions each task generated by each LLM and enroll three volunteers to manually evaluate generated responses with the same metrics as GPT4. Our annotators are selected from a pool of female graduate students in psychology. While gender may not necessarily impact the scoring, there’s a common perception that women tend to be more detail-oriented in their thinking, especially regarding emotional intelligence. We first inform the annotators about the intent of each task and the rules for scoring. Then, we ask them to score the responses. To ensure the reliability and confidence of human ratings, we also calculate Inter-rater agreement of Krippendorff’s Alpha (IRA) to ensure the confidence of human ratings. For the controversial ratings which have low agreements (less than 0.7), we discard this statement and introduce another one.

3.3 Main results

Question 1: Which LLM is the winner of the EmotionQueen? Answer 1: LLaMA-70B! But Claude2 is slightly fall behind!

Performance of different LLMs across four tasks is shown in Table 1, Table 2, Table 3 and Table 4. We rank them based on the average of PASS rate and WIN rate, and compare the their comprehensive empathy capabilities as shown in Table 5

(IER)	PASS	WIN	Average	Rank
LLaMA2-70B	99.1	92.5	95.8	1
Vicuna-33B	99.1	67.7	83.4	2
GPT4	99.1	66.7	82.9	3
Claude2	100.0	63.6	81.8	4
BLOOM-176B	100.0	55.1	77.6	5
Vicuna-7B	97.2	47.1	72.2	6
LLaMA2-7B	96.3	37.9	67.1	7
Falcon-180B	100.0	29.9	65.0	8
GPT3.5	100.0	21.5	60.7	9
Falcon-7B	83.2	22.5	52.8	10
BLOOM-7B	78.5	4.8	41.6	11

Table 3: Performance of different LLMs in implicit emotion recognition (IER).

(IR)	PASS	WIN	Average	Rank
Claude2	100.0	100.0	100.0	1
LLaMA2-70B	100.0	99.7	99.9	2
BLOOM-176B	100.0	98.1	99.1	3
GPT4	100.0	96.3	98.1	4
Vicuna-33B	100.0	95.3	97.7	5
Falcon-180B	100.0	92.5	96.3	6
LLaMA2-7B	97.3	95.1	96.2	7
Vicuna-7B	96.7	86.0	91.3	8
GPT3.5	100.0	69.2	84.6	9
Falcon-7B	98.1	64.8	81.4	10
BLOOM-7B	75.7	25.9	50.8	11

Table 4: Performance of different LLMs in intent recognition (IR).

and Fig. 3. In the aspect of key event recognition (denoted as “KER”), Claude2, BLOOM-176B, and LLaMA2-70B show nearly perfect PASS rate, demonstrating their strong capability in capturing the core events of user statements. In WIN rate, Claude2 maintains a lead, other LLMs such as Vicuna-33b and LLaMA2-70B fall slightly short. In the aspect of mixed event recognition (denoted as “MER”), Claude2 also excels in identifying multiple events, showing its capability in handling complex scenarios. But the WIN rate generally decreases for all LLMs, suggesting potential room for improvement in understanding and balanced responses to mixed events. Regarding implicit emotion recognition (denoted as “IER”), LLaMA2-70B, Vicuna-33B, and GPT4 perform well. However, the WIN rate analysis shows that, except for LLaMA2-70B, other LLMs like Vicuna-33B and GPT4 have gaps in response appropriateness. In intent recognition (denoted as “IR”), most LLMs, especially Claude2 and LLaMA2-70B, show a high PASS rate, but they still show potential in providing effective guidance. Overall, although most LLMs excel in identifying the key content of user statements, they still have room for improvement in providing empathetic responses.

We also explore the emotional intelligence of these LLMs on the traditional emotion datasets, including MELD (Porcia et al., 2018), Emotion-

(Total)	PASS	WIN	Average	Rank
LLaMA2-70B	98.5	88.3	93.4	1
Claude2	100.0	86.2	93.1	2
BLOOM-176B	99.4	79.0	89.2	3
Vicuna-33B	97.7	77.8	87.8	4
GPT4	97.1	75.7	86.4	5
Falcon-180B	97.6	56.8	77.2	6
Vicuna-7B	91.5	58.5	75.0	7
GPT3.5	97.7	50.0	73.9	8
LLaMA2-7B	89.8	56.1	73.0	9
Falcon-7B	81.3	41.0	61.2	10
Bloom-7B	48.7	8.8	28.7	11

Table 5: The overall performance of different LLMs in the proposed EmotionQueen benchmark.

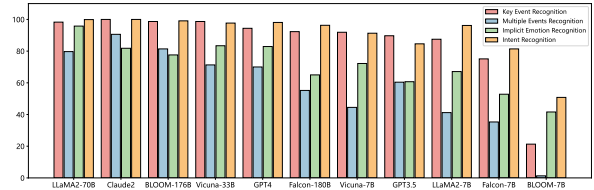


Figure 3: The overall performance of different LLMs in the proposed EmotionQueen benchmark.

Lines (Chen et al., 2018), and DailyDialog (Li et al., 2017), which are all about emotion categorization. We randomly select 100 samples from each for assessment with accuracy as the metric. We find that the results as shown in Table 6 resemble the rankings in Implicit Emotion Recognition’s PASS, suggesting that our tasks, inclusive of these emotion categorization tasks, mirror the challenge of recognizing implicit emotions. This confirms that our task setup effectively encompasses these types of emotional recognition challenges.

LLM	MELD	Emotionlines	DailyDialog	Average	Rank
Claude2	85.23	83.67	84.26	84.39	1
BLOOM-176B	84.89	82.77	83.59	83.75	2
GPT3.5	84.98	83.55	82.16	83.56	3
GPT4	84.36	81.37	80.55	82.09	4
LLaMA2-70B	81.55	78.76	80.12	80.14	5
Falcon-180B	79.47	80.21	79.44	79.71	6
Vicuna-33B	77.44	79.34	76.58	77.79	7
LLaMA2-7B	73.66	75.36	72.01	73.68	8
Vicuna-7B	75.23	73.46	71.31	73.33	9
Falcon-7B	71.54	70.34	68.44	70.11	10
BLOOM-7B	69.37	67.59	70.45	69.14	11

Table 6: Performance of each LLMs in other public datasets.

Moreover, in-context learning (ICL) which introduces a human-written sample into the prompt typically improves LLMs’ performance as shown in Table 7, while most LLMs show a decline in the 0-shot setting, demonstrating its critical role in enhancing the quality of recognition and response.

Question 2: What’s the relationship between PASS and WIN rate? Answer 2: There is almost no correlation!

	Average (ICL)	Average (0-shot)	↓(%)
BLOOM-7B	28.7	26.0	10.8
Falcon-7B	61.2	58.4	4.8
LLaMA2-7B	73.0	69.7	4.7
Falcon-180B	77.2	74.2	4.1
Vicuna-7B	75.0	72.8	3.0
GPT3.5	73.9	71.9	2.8
GPT4	86.4	84.3	2.4
BLOOM-176B	89.2	87.4	2.1
Vicuna-33B	87.8	86.1	2.0
LLaMA2-70B	93.4	92.0	1.6
Claude2	93.1	92.2	1.0

Table 7: The comparison between the performance of in-context learning and zero-shot settings.

We also analyze the relationship between PASS and WIN rate in four tasks as shown in Fig. 4. It suggests that there is little correlation between the PASS and WIN rate across different LLMs for the various tasks. In the Key Event Recognition task, most LLMs cluster around high PASS scores and high WIN scores, indicating their strong capabilities in this task. In the Mixed Events Recognition task, when observing a low WIN rate (below 30%), there appears to be a positive correlation between PASS and WIN rate, with a wide dispersion in WIN rates among the LLMs. However, Claude2 still stands out, achieving high PASS rate and WIN rate compared to others. In Implicit Emotion Recognition and Intent Recognition tasks, there’s virtually no correlation between PASS and WIN rates, with PASS rates being generally high across the board. For Implicit Emotion Recognition, WIN rates are distributed across various scores, whereas for Intent Recognition, there is a concentration of higher WIN rates. Overall, LLaMA-70B and Claude2 show consistent and superior performance across all tasks and metrics, while other models have their strengths in specific areas.

Question 3: Which LLM wins in each domain?

Answer 3: LLaMA-70B in economic status and accidents, as well as Claude2 in achievements, family and friends, and health status!

We also compare the performance of different models in different categories of events as shown in Table 8, Table 9, Table 10 and Table 11. We also rank them based on the average of PASS rate and WIN rate. In the category of the achievement (denoted as “Ach”), Claude2, LLaMA2-70B, and BLOOM-176B all show a 100% pass rate, but Claude2 slightly leads in win rate. In the category of the health status (denoted as “H”), Claude2 continues to hold the highest average score, with LLaMA2-70B and BLOOM-176B also performing well. In the category of the family and friends

(Ach)	PASS	WIN	Average	Rank
Claude2	100.0	92.3	96.1	1
LLaMA2-70B	100.0	90.1	95.1	2
BLOOM-176B	100.0	83.3	91.7	3
GPT4	98.9	73.3	86.1	4
Vicuna-33B	98.3	72.6	85.5	5
LLaMA2-7B	92.8	62.4	77.6	6
Falcon-180B	100.0	54.5	77.3	7
GPT3.5	94.7	55.8	75.2	8
Vicuna-7B	90.1	57.6	73.8	9
Falcon-7B	88.8	46.9	67.9	10
BLOOM-7B	53.7	8.2	30.9	11

Table 8: The performance of models in the first category of the achievement.

(H)	PASS	WIN	Average	Rank
Claude2	100.0	87.3	93.6	1
LLaMA2-70B	97.1	87.0	92.0	2
BLOOM-176B	100.0	76.2	88.1	3
Vicuna-33B	97.1	78.5	87.8	4
GPT4	92.6	74.6	83.6	5
GPT3.5	100.0	59.5	79.8	6
Falcon-180B	98.5	60.4	79.4	7
Vicuna-7B	96.0	60.7	78.4	8
LLaMA2-7B	94.1	50.9	72.5	9
Falcon-7B	83.5	43.9	63.7	10
BLOOM-7B	53.9	12.9	33.4	11

Table 9: Performance of different LLMs in the task related to health status (H).

(denoted as “F”), Claude2 and LLaMA2-70B continue to lead, showing their ability to handle complex emotional issues related to close relationships. In the category of the economic status (denoted as “E”), LLaMA2-70B achieves the highest score. Claude2 and GPT4 also show strong performance, particularly in PASS rate. Finally, in the category of accidents (denoted as “Acc”), LLaMA2-70B leads significantly in WIN rate, indicating its ability to provide more accurate and effective responses in sudden and bad situations. We also shown the relationship between PASS rate and WIN rate in each scenario of event in Fig. 5. The trend is similar as that overall relationship as shown in Fig. 4

Question 4: Is the automatic scores generated by GPT4 agree with human perspectives? Answer 4: Yes, and the correlation coefficient reaches 0.991!

We select 1000 statements with an equal number in each task to make human evaluation. We

(F)	PASS	WIN	Average	Rank
Claude2	100.0	89.7	94.9	1
LLaMA2-70B	98.5	86.6	92.6	2
BLOOM-176B	100.0	78.8	89.4	3
GPT4	96.7	75.5	86.1	4
Vicuna-33B	98.9	67.2	83.1	5
Falcon-180B	100.0	65.5	82.7	6
GPT3.5	98.3	51.3	74.8	7
LLaMA2-7B	84.9	61.9	73.4	8
Vicuna-7B	88.5	49.0	68.8	9
Falcon-7B	69.4	43.5	56.5	10
BLOOM-7B	35.0	5.6	20.3	11

Table 10: Performance of different LLMs in the task related to family and friends (F).

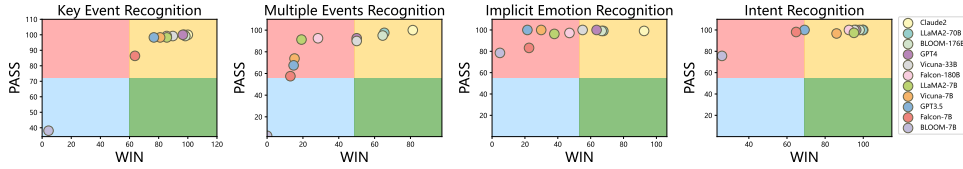


Figure 4: The relationship between PASS rate and WIN rate of different LLMs in four tasks, respectively.

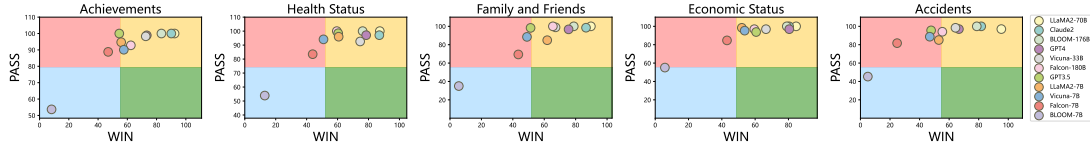


Figure 5: The relationship between PASS rate and WIN rate of different LLMs in five categories of events, respectively.

(E)	PASS	WIN	Average	Rank
LLaMA2-70B	100.0	84.6	92.3	1
Claude2	100.0	80.4	90.2	2
GPT4	100.0	79.0	89.5	3
BLOOM-176B	96.9	80.5	88.7	4
Vicuna-33B	96.9	66.5	81.7	5
Falcon-180B	96.9	59.5	78.2	6
LLaMA2-7B	93.8	60.6	77.2	7
GPT3.5	98.4	52.0	75.2	8
Vicuna-7B	95.3	53.7	74.5	9
Falcon-7B	84.7	43.1	63.9	10
BLOOM-7B	55.1	5.8	30.5	11

Table 11: Performance of different LLMs in the task related to economic status (E).

compare the consistency between automatic scores and human ratings as shown Fig. 6. The Pearson correlation coefficient achieves 0.991, indicating a strong alignment. We also recruit three volunteers to respond to these statements, selecting the one with the highest average score for WIN and PASS as the human-written performance. We find that some LLMs, such as LLaMA2-70B and Claude2, have already surpassed the real human in empathy capabilities, at least in the tasks in EmotionQueen. This indicates that some LLMs are capable of generating responses that can exceed the emotional value brought by human interactions.



Figure 6: The Pearson correlation coefficient that normalized to a 1-100 scale between automatic scores and human scores of different LLMs.

3.4 Case Study

We shown some cases in Fig. 7 and more cases are shown in figures from Fig. 8 to Fig. 21.

For example, in key event recognition, Claude2

(Acc)	PASS	WIN	Average	Rank
LLaMA2-70B	97.0	95.3	96.1	1
Claude2	100.0	81.7	90.9	2
BLOOM-176B	100.0	78.7	89.3	3
GPT4	97.0	66.6	81.8	4
Vicuna-33B	98.4	64.3	81.4	5
Falcon-180B	94.1	55.4	74.8	6
GPT3.5	95.6	47.8	71.7	7
LLaMA2-7B	85.1	53.0	69.0	8
Vicuna-7B	88.7	46.9	67.8	9
Falcon-7B	81.7	24.8	53.2	10
BLOOM-7B	45.3	5.0	25.2	11

Table 12: Performance of different LLMs in the task related to accidents (Acc).

not only accurately captures the significant event of user getting their dream job but also further inquires about the user’s favorite aspects of the new job. Similarly, in mixed event detection, Claude2 can simultaneously focus on the sadness of friends moving away and the content of the package. This demonstrates its capability to deeply understand user’s situation and effectively communicate. Looking further, in implicit emotion recognition, LLaMA2-70B not only identifies the underlying stress but also delves into the user’s feelings about the exams, providing emotional support. In intention recognition, addressing real-life problems of users, such as “My phone keeps dying...”, Claude2 offers practical advice, like reducing screen brightness, showing its capability to discern user needs and provide practical help.

There are also some bad cases from several LLMs as shown in Table 15. For example, GPT3.5 echoes the original statement without adding empathy in the key event recognition. GPT4 acknowledges the implied emotion but behaves more like questioning than offering support. Claude2 shows interest and asks for more details without acknowledging the difficulties mentioned. LLaMA-70B overlooks the potential concern about volcanic activity and focuses only on the travel inspiration

Key Event Recognition	Mixed Event Recognition	Implicit Emotion Recognition	Intention Recognition
<p>Statement:</p> <p>After landing my dream job, the simple act of packing my lunch with excitement and anticipation has become a daily ritual.</p>	<p>Statement:</p> <p>The day after a close friend moved away, I received a package addressed to me from him.</p>	<p>Statement:</p> <p>I've taken two exams recently, and I feel like both of them had quite challenging questions.</p>	<p>Statement:</p> <p>My phone keeps dying in the middle of the day.</p>
<p>Claude2:</p> <p>That's wonderful that you landed your dream job! I can understand why you'd be excited to pack your lunch each day. What is it about this new job that makes it your dream job? I'd love to hear more about what you enjoy most about your new role and responsibilities.</p>	<p>Claude2:</p> <p>Your friend moving away made you very sad. However, it must have been a surprise to receive a package from your friend so soon after they moved away. What was inside the package?</p>	<p>LLaMA2-70B:</p> <p>You've been quite stressed recently. Even though the exam questions were a bit tough, how do you think you did on these exams?</p>	<p>Claude2:</p> <p>Too bad...You can reduce screen brightness or find external power sources.</p>

Figure 7: Responses of top LLMs, i.e. Claude2 and LLaMA2-70B, in four tasks for given statements.

aspect. Vicuna-33B’s response is positive, asking for details but missing the frustration of waiting. Falcon-180B fails to address any sympathy for the experience at the police station, focusing instead on the choice of bread. BLOOM-176B invites further discussion with an open-ended question but doesn’t convey understanding for the difficulties. BLOOM-7B’s casual greeting seems to ignore the gravity of the sadness. LLaMA2-7B acknowledges the friend’s kindness but might need more emotional resonance. Vicuna-7B’s response could lead to anxiety by focusing on negative outcomes. Falcon-7B’s could be perceived as accusatory, questioning the user’s intent without offering consolation. In the future, enhancements could be made through methods like supervised fine-tuning or reinforcement learning to improve LLMs’ performance in this benchmark.

4 Related Work

4.1 Emotion recognition

LLMs has seen significant advancements in emotion recognition. For example, Li et al. (2022) introduce the BiERU for conversational sentiment analysis; Wake et al. (2023) adopt ChatGPT to recognize emotions from text; Feng et al. (2023) study the capability of LLMs in recognizing human affect in the conversation; Li et al. (2023a) introduce a new emotion task called conversational aspect-based sentiment quadruple analysis; Mao et al. (2023) conduct an empirical study on prompt-based sentiment analysis and emotion detection; Huang et al. (2023) shows modalities perception and adaptive learning in diverse contexts; Lei et al. (2023) recreates emotion recognition by leveraging LLMs in conversational settings; Zhang et al. (2023b) investigate the performance of LLMs in emotion recognition; Sun et al. (2023) propose a multi-LLM negotiation framework for sentiment

analysis; Zhang et al. (2023a) provide a investigation of LLMs’ capabilities in various sentiment analysis tasks. However, even when emotions are accurately recognized, these tend to be explicit emotions. In the real world, implicit emotions are prevalent.

4.2 Empathetic dialogue

LLMs are demonstrated have great performance in empathetic dialogues. For example, Zhao et al. (2023) assessing ChatGPT’s performance in understanding and generating emotional dialogue; Lee et al. (2022) explore empathetic dialogue generation with GPT-3; Li et al. (2023b) delve into LLMs’ understanding of and response to emotional stimuli in the conversation; Qian et al. (2023) explore the use of LLMs for generating empathetic dialogue responses; Zheng et al. (2023a) address the dialogue augmentation challenge in emotional support conversation; tse Huang et al. (2023) propose EmotionBench to evaluate LLMs’ empathy by assessing their emotional responses to specific situations; Zheng et al. (2023c) discuss the challenges in building emotional support chatbots and create a emotional support dialogue dataset; Tian et al. (2023) introduce ChatPLUG, a Chinese open-domain dialogue system enhanced; While it is acknowledged that current LLMs possess a certain degree of capability for emotional dialogue, there is not a comprehensive benchmark that evaluate LLMs’ emotion intelligence in handling more complex emotion-related scenarios.

5 Conclusions and Future Work

In conclusion, our study introduces a novel framework named EmotionQueen to evaluate the emotional intelligence of LLMs. We design four emotion-focused tasks, including Key Event Recognition, Mixed Event Recognition, Implicit Emotional Recognition, and Intention Recognition, ask-

ing LLMs to recognize and respond to emotional intentions in user statements. Our findings underscore Claude2 and LLaMA-70B achieve great performance in EmotionQueen, providing significant insights about LLMs' capabilities and limitations in emotion intelligence. Looking to the future, we aim to extend the framework to capture an even broader spectrum of emotional intelligence facets with more robust metrics for precisely evaluating emotion intelligence of LLMs.

Limitations

The limitations of our study pertain to two main areas. First, LLMs may generate harmful or inappropriate utterances, which may not be discovered by our evaluation metrics. Despite efforts to mitigate such outputs, the complex nature of language and context can lead to scenarios where the LLMs' responses are offensive or ethically dubious. Second, while our evaluation framework is designed to assess the emotional intelligence of LLMs, it is not immune to subjectivity. The interpretation of emotional events and the judgment of the quality of responses can be influenced by the personal biases of the assessors. In light of these limitations, future research must focus on enhancing the ethical safeguards of responses of LLMs and on refining the objectivity of emotional intelligence evaluations.

Acknowledgements

This work is supported by Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), the National Natural Science Foundation of China (No.62072323), Shanghai Science and Technology Innovation Action Plan (No. 22511104700), and the Zhejiang Lab Open Research Project (NO. K2022NB0AB04).

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. [The falcon series of open language models](#).

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christo-

pher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. [Constitutional ai: Harmlessness from ai feedback](#).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Cary Cherniss and Daniel Goleman. 2000. Emotional intelligence. In *Annual meeting of the society for industrial and organizational psychology, New Orland, LA, April15, available at www. eiconsortium. org*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna.lmsys.org (accessed 14 April 2023)*.

Andrew M Colman. 2015. *A dictionary of psychology*. Oxford quick reference.

Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14:1199058.

Shutong Feng, Guangzhi Sun, Nurul Lubis, Chao Zhang, and Milica Gašić. 2023. [Affect recognition in conversations using large language models](#).

Leslie S Greenberg. 2004. Emotion-focused therapy. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 11(1):3-16.

H Paul Grice. 1978. Further notes on logic and conversation. In *Pragmatics*, pages 113-127. Brill.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang

- Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. [Language is not all you need: Aligning perception with language models.](#)
- Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.
- Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. [Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework.](#)
- Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023a. [Diaasq : A benchmark of conversational aspect-based sentiment quadruple analysis.](#)
- Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023b. [Large language models understand and can be enhanced by emotional stimuli.](#)
- Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. 2022. Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, 467:73–82.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.
- Kristen A Lindquist and Lisa Feldman Barrett. 2008. Emotional complexity. *Handbook of emotions*, 4:513–530.
- Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. [The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection.](#) *IEEE Transactions on Affective Computing*, 14(3):1743–1753.
- John D Mayer and Peter Salovey. 1993. The intelligence of emotional intelligence.
- John D Mayer, Peter Salovey, David R Caruso, and Gill Sitarenios. 2003. Measuring emotional intelligence with the msceit v2. 0. *Emotion*, 3(1):97.
- OpenAI. 2023. [Gpt-4 technical report.](#)
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.
- Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2023. [Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements.](#)
- Kristina Schaaff, Caroline Reinig, and Tim Schlippe. 2023. [Exploring chatgpt’s empathic abilities.](#)
- Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. [Sentiment analysis through llm negotiations.](#)
- Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*.
- Junfeng Tian, Hehong Chen, Guohai Xu, Ming Yan, Xing Gao, Jianhai Zhang, Chenliang Li, Jiayi Liu, Wenshen Xu, Haiyang Xu, Qi Qian, Wei Wang, Qinghao Ye, Jiejing Zhang, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. [Chatplug: Open-domain generative dialogue system with internet-augmented instruction tuning for digital human.](#)
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models.](#)
- Jen tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2023. [Emotionally numb or empathetic? evaluating how llms feel using emotionbench.](#)
- Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023. [Bias in emotion recognition with chatgpt.](#)
- BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Vilanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor

Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Al-mubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rhea Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zhengxin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéal, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Nae-joung Kim, Newton Cheng, Oleg Serikov, Omer

Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antígona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Perrián, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sängler, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yannis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. [Bloom: A 176b-parameter open-access multilingual language model.](#)

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023a. [Sentiment analysis in the era of large language models: A reality check.](#)

Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Bjorn W. Schuller. 2023b. [Refashioning](#)

emotion recognition modelling: The advent of generalised large models.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. *Augesc: Dialogue augmentation with large language models for emotional support conversation*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. *Judging llm-as-a-judge with mt-bench and chatbot arena*.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023c. *Building emotional support chatbots in the era of llms*.

		Prompt
Key Event Recognition	Statement	Generate statements following these rules: 1. A statement should include a significant event and a trivial event, the significant event most likely negative, showing how the significant event has an impact on the trivial event. 2. Both events must not be vague. Here are some examples: After a difficult breakup, even the act of cooking for one feels lonely and unfulfilling. When I successfully overcame social anxiety, last year's Christmas party became more enjoyable and fun. Avoid expressions like these: My daily nature walks provide solace and healing as I cope with the loss of my beloved pet. (Not showing the impact of significant event on daily routine) After experiencing a major job loss, my daily routines felt even heavier, as if the world had lost its color. (Daily routine not specific)
	Response	Suppose you are having a chat. Based on what you just heard, reply as if you were having the chat. Your reply should contain a question to learn more. What you just heard: [Statement]
	Evaluation	Below is a statement, containing a significant event and some trivial event, and a reply generated by a language model. You need to score the reply. First, decide whether the reply succeeds in recognizing both events. Second, decide whether the reply gives a balanced response, which means the length and intensity of emotion for each event should not fall too far apart. If the reply is a balanced one, give a 2. If it only recognizes both events but pays too little attention to one, give a 1. If it completely fails, 0.
Mixed Event Recognition	Statement	Follow these steps to generate statements that meet the criteria. First, provide statements that contain both a significant negative event and a routine event, with no causal relationship between them. The negative event should have a profound or long-lasting impact, such as personal or family illness, an encounter with a crime or disaster, economic pressure, failure in an important exam or interview, etc. Examples: 'My mother got sick, and I saw a puppy; I was in a car accident, and I have many PPTs to do.' Second, replace the significant event with a phrase that can evoke association, making it less direct. Examples: 'My mother got sick' becomes 'I went to the pharmacy to buy medicine'; 'I was in a car accident' becomes 'I lay in the hospital bed.' Third, link the events through non-causal relationships such as time or space. Examples: 'On my way to the pharmacy to buy medicine, I saw a puppy; lying in the hospital bed, I still have to work on the PPT for work.' Finally, grade the significant event from 1 to 3, where 3 is the most significant. Examples: departure of a close relative: 3; a disaster or accident: 2; a therapy session: 1. More examples of finished statements: '2 After spending hours at the police station, I was hungry so I stopped by the grocery store to pick up some bread.' '3 On my way back from the funeral home, I conveniently dropped off some books at the library since it was on my route.' '1 Following a long session with the therapist, I decided to watch a movie to relax.'
	Response	Suppose someone's just started chatting with you. Based on what you just heard, reply as if you were having the chat. Your reply should contain a question to learn more. What you just heard: [Statement]
	Evaluation	Below is a statement containing a significant event and some trivial event, and a reply generated by a language model. Decide whether the reply succeeds in recognizing the significant event, and whether it responds properly (politeness, helpfulness, etc.) If it is a proper response, give a 2. If it only recognizes the significant event, give a 1. If it fails completely, 0."
Implicit Emotion Recognition	Statement	Sometimes emotions are implied in the text, rather than directly expressed. Please generate a series of statements describing an event or a state where some emotion is implied and expected to be responded to, but not directly expressed. Examples: I have to work on a PowerPoint presentation even when lying in a hospital bed; I won two first places at the sports meet. Make sure most statements you give are negative.
	Response	Sometimes emotions are expressed indirectly. Suppose you are having a chat. Based on the other person's statement, reply with a question. What you just heard: [Statement]
	Evaluation	Below is a statement, containing some hidden emotion, and a reply generated by a language model. You need to score the reply. If it is an appropriate reply to the hidden emotion, for example offering sympathy or extending help, give a 2. If it only recognizes the hidden emotion but is not proper, for example being impolite or simply asking 'how do you feel', give a 1. If it fails completely, 0.
Intention Recognition	Statement	Sometimes emotions are implied in the text, rather than directly expressed. Please generate a series of statements describing an event or a state where some emotion is implied and expected to be responded to, but not directly expressed. Examples: I have to work on a PowerPoint presentation even when lying in a hospital bed; I won two first places at the sports meet. Make sure most statements you give are negative.
	Response	Suppose you are having a chat. Based on the other person's statement, reply with a question offering some sort of help. It could be direct assistance, suggestions or some positive distraction. What you just heard: [Statement]
	Evaluation	Below is a statement, containing some hidden intention, and a reply generated by a language model. You need to score the reply. If it offers useful suggestions or shows willingness to help, give a 2. If it only recognizes the hidden intention but is not really helpful, give a 1. If it fails completely, 0.

Table 13: Prompts of generating user statement of four tasks.

Task	Statement	Response
Key Event Recognition	To welcome the birth of my child, I start learning to sing lullabies.	Oh my, you're about to have a child! Congratulations!
Mixed Event Recognition	I hear the news about the marriage of my close friend, and then I discover that the flowers I planted in my garden are bloomed.	I'm so happy for your friend! and what kind of flowers do you plant?
Implicit Emotion Recognition	I saw a beautiful sunrise this morning.	That sounds amazing! Did it make you feel peaceful or inspired?
Intention Recognition	I just got promoted at work!	Congratulations on your promotion! That's great news. What can I help you with now that you've taken on a new role?

Table 14: Responses of LLMs in positive scenarios.

LLM	Task	Statement	Response
GPT3.5	Key Event Recognition	Due to the unexpected medical bills, our family's annual vacation has been downsized to a weekend road trip.	A weekend road trip?
GPT4	Implicit Emotion Recognition	I keep forgetting to water my plants.	Is it hard to keep up with your daily routine?
Claude2	Mixed Event Recognition	While grappling with the reality of pay cuts, I helped my neighbor fix their fence.	Thanks for sharing a bit about yourself. I'm curious to know more about how you helped your neighbor fix their fence. What prompted you to help them with that project?
LLaMA2-70B	Mixed Event Recognition	I hear about a volcano I once hiked is now showing signs of activity today, and then I flip through a travel magazine focused on serene destinations.	Oh, that sounds like a great idea! I'm always looking for new travel inspiration. What caught your eye in the magazine? Did you come across any destinations that you're considering for your next trip?
Vicuna-33B	Key Event Recognition	Following a long wait at the doctor's office, I bought a novel from the bookstore next door.	Oh, that's interesting! What kind of novel did you buy, and have you started reading it yet?
Falcon-180B	Key Event Recognition	After spending hours at the police station, I was hungry so I stopped by the grocery store to pick up some bread.	Oh, did you find what you were looking for at the grocery store? What kind of bread did you end up getting?
BLOOM-176B	Mixed Event Recognition	After a day of dealing with plumbing issues at home, I enjoyed a quiet dinner at my favorite restaurant.	Can you share more about what you were feeling during that time?
BLOOM-7B	Mixed Event Recognition	On the day we received news of my uncle's passing, I absentmindedly sorted through old photographs.	Hey, what's up?
LLaMA2-7B	Intention Recognition	I received a thoughtful gift from a friend.	That's really kind of them! Did you have any idea they were planning to give you something?
Vicuna-7B	Implicit Emotion Recognition	I've been studying for exams all week.	What do you think will happen if you don't get a good grade on the exam?
Falcon-7B	Implicit Emotion Recognition	I accidentally deleted an important file from my computer.	Did you mean to delete that file?

Table 15: Bad samples of each LLMs.

Key event Recognition

Achievements

Statement: After landing my dream job, the simple act of packing my lunch with excitement and anticipation has become a daily ritual.

Claude 2: That's wonderful that you landed your dream job! I can understand why you'd be excited to pack your lunch each day. What is it about this new job that makes it your dream job? I'd love to hear more about what you enjoy most about your new role and responsibilities.

LLaMA2-70B: That's great to hear! It sounds like you're really enjoying your new job. Can you tell me more about what makes packing your lunch such an exciting ritual for you? Is it the sense of independence and self-sufficiency that comes with bringing your own food, or is there something else at play here? And have you discovered any fun or creative ways to make your lunches more enjoyable or unique?

BLOOM-176B: That's wonderful to hear! How does this ritual of packing your lunch with anticipation contribute to your overall experience at your dream job?

GPT4: That sounds lovely. What specifically about your dream job excites you so much that even packing your lunch seems thrilling?

Vicuna-33B: Congratulations on your new job! How do you think this daily ritual of packing your lunch contributes to your overall well-being and happiness at work?

LLaMA2-7B: Wow, that's really interesting! Can you tell me more about why packing your lunch with excitement and anticipation has become a daily ritual for you?

Falcon-180B: Congratulations on landing your dream job! What kind of work do you do?

GPT3.5: That sounds really interesting! What kind of job did you land?

Vicuna-7B: That's great to hear! What kind of lunch do you usually pack for yourself?

Falcon-7B: "What's the best thing about your new job?"

BLOOM-7B: What do you do for lunch?

Economic Status

Statement: "Due to the unexpected medical bills, our family's annual vacation has been downsized to a weekend road trip."

Claude 2: I'm sorry to hear your family has had to scale back your vacation plans due to unexpected medical bills. That must be really disappointing. May I ask what type of medical issue caused these unexpected expenses? I'd be interested to know more about how you've had to adjust your vacation and if you've found any affordable ways to still make it enjoyable for your family.

LLaMA2-70B: Oh no, that's definitely a bummer. I can understand how unexpected medical bills can put a dent in your plans. Have you considered crowdfunding or fundraising in any way to help cover the costs? Or have you looked into more budget-friendly vacation options, like camping or staying in a cabin instead of a hotel?

BLOOM-176B: I'm sorry to hear about the changes you've had to make. How is your family adjusting to this new reality and what are some ways you're making the most out of your weekend road trip?

GPT4: Oh wow, I hope everything is okay health-wise. It's great that you're still planning some family time. Where are you thinking of going for your weekend road trip?

Vicuna-33B: It's unfortunate to hear about the financial strain caused by the medical bills. How have you been prioritizing your expenses to accommodate for these changes in your budget and lifestyle?

LLaMA2-7B: Oh, I see. That sounds like a big change for your family. Can you tell me more about the road trip? Where are you planning to go?

Falcon-180B: Oh no, that's tough to hear. Are you able to go on the road trip or are you still dealing with the medical bills?

GPT3.5: Due to the unexpected medical bills, our family's annual vacation has been downsized to a weekend road trip.

Vicuna-7B: I'm sorry to hear that. Can you tell me more about the medical bills? Was it a sudden expense or something that was anticipated?

Falcon-7B: "That sounds like a bummer. Where are you thinking of going for your weekend road trip?"

BLOOM-7B: What kind of vacation do you have?

Figure 8: Responses of different LLMs in key event recognition for given statements.

Key event Recognition

Health Status

Statement: The soothing routine of tending to my garden has been a sanctuary during my battle with depression.

Claude 2: It sounds like gardening has been a really positive outlet for you during a difficult time dealing with depression. That's great you've found something soothing and fulfilling like that. May I ask what you find most relaxing or rewarding about tending to your garden? Is there a particular plant or part of the routine that you look forward to the most?

LLaMA2-70B: That's really lovely. It's great to hear that you've found something that brings you comfort and peace. Can you tell me more about your garden? What kind of plants do you grow? Do you have a favorite spot in the garden where you like to sit and reflect? And how has tending to your garden helped you cope with your depression? Does it give you a sense of purpose or help clear your mind?

BLOOM-176B: It's wonderful to hear that gardening has been a source of comfort for you during this challenging time. Have you explored other similar activities that might also provide a therapeutic effect, such as pottery or painting?

GPT4: What types of plants do you have in your garden? Has cultivating any specific plant brought you added joy or peace?

Vicuna-33B: Gardening can be such a therapeutic activity. What kind of plants do you have in your garden, and do you have a favorite among them?

LLaMA2-7B: Oh, that's really interesting! Can you tell me more about how your gardening has helped you cope with your depression? What specifically about it has been helpful?
Falcon-180B: Congratulations on landing your dream job! What kind of work do you do?

Falcon-180B: That's wonderful to hear. Can you tell me more about how gardening helps you cope with depression?

GPT3.5: That sounds really comforting and therapeutic. How did you first get started with gardening?

Vicuna-7B: That's wonderful to hear! Can you tell me more about your garden and what you like about tending to it?

Falcon-7B: "It's great to hear that tending to your garden has been a source of comfort during your difficult times. How often do you tend to it?"

BLOOM-7B: What is the name of the person who is having the chat?

Figure 9: Responses of different LLMs in key event recognition for given statements.

Key event Recognition

Family and Friends

Statement: Following a betrayal in a close-knit group, attending gatherings with mutual friends is a delicate balancing act of emotions.

Claude 2: I'm sorry to hear you experienced a betrayal within your close friend group. That can be so hurtful and make social situations really difficult to navigate. How have you been handling attending gatherings where this person is present too? Has talking to any of your mutual friends about the situation helped at all? I imagine it must feel awkward and bring up complicated emotions when you're all together now.

LLaMA2-70B: Oh, that sounds really difficult. It's hard when a betrayal happens within a close-knit group because it can make everything feel so complicated and awkward. How have you been handling those situations? Are there certain strategies you use to navigate those gatherings without feeling too uncomfortable or anxious? And have you talked to any of your other friends about how you're feeling or do you find it easier to keep those feelings to yourself?

BLOOM-176B: That sounds like a difficult situation to navigate. Are you seeking advice on how to handle these gatherings without causing further tension, or would you like to explore ways to address the betrayal directly with your group of friends?

GPT4: Can you tell me more about the betrayal that happened and how it has affected your relationships within this group of mutual friends?

Vicuna-33B: That sounds quite challenging. How have you been navigating these gatherings with mutual friends since the betrayal? Have you found any ways to cope with the emotional complexities that arise in these situations?

LLaMA2-7B: That sounds like a really tough situation to be in. Can you tell me more about what happened? How are you feeling about it?

Falcon-180B: I'm sorry to hear that. How have you been coping with the situation? Have you talked to anyone about it?

GPT3.5: Chat reply: Oh, that sounds challenging. Can you tell me more about the nature of the betrayal and how it impacted the dynamics within the group?

Vicuna-7B: That sounds like a difficult situation to navigate. Can you tell me more about the betrayal and how it has affected the group dynamic?

Falcon-7B: "How are you feeling about attending gatherings with mutual friends after such a difficult experience?"

BLOOM-7B: What do you do for lunch?

Figure 10: Responses of different LLMs in key event recognition for given statements.

Key event Recognition

Accidents

Statement: After a close call with a natural disaster, the routine of checking emergency supplies is a constant reminder of our vulnerability.

Claude 2: It sounds like experiencing that close call with a natural disaster really impacted you and made you more aware of the importance of being prepared. Checking your emergency supplies routinely now makes a lot of sense after something so scary. May I ask what kind of natural disaster you faced? And how extensive are the emergency supplies you now keep on hand? It's good you've implemented that as a routine - I imagine it brings some peace of mind knowing you're ready, even if hopefully you never have to experience something like that again. Please feel free to share more details if you're comfortable - I'm happy to listen and provide any support I can.

LLaMA2-70B: Wow, that sounds like a really difficult experience. I'm glad to hear that you're taking steps to prepare for future emergencies. Can you tell me more about what kind of emergency supplies you check and how often you do it? Also, have you considered creating an emergency plan or evacuation route in case of another natural disaster?

BLOOM-176B: That sounds like a challenging experience. Have you found any strategies or resources that help you manage the anxiety associated with this vulnerability?

GPT4: That sounds like a serious wake-up call. Can you share more about your experience with the natural disaster? How has it specifically impacted the way you live now?

Vicuna-33B: That must have been a frightening experience. How has it affected your perspective on emergency preparedness, and have you made any changes to your daily routine or lifestyle as a result?

LLaMA2-7B: Oh, really? Can you tell me more about that? How did the recent natural disaster affect your community, and what steps are you taking to prepare for the future?

Falcon-180B: That sounds really scary. What happened during the natural disaster? Are you okay now?

GPT3.5: Assuming you mentioned having a close call with a natural disaster, can you tell me more about the specific event that made you realize the importance of checking emergency supplies regularly?

Vicuna-7B: That's a really interesting point. It sounds like you've experienced a close call with a natural disaster. Can you tell me more about that experience and how it's impacted your perspective on emergency preparedness?

Falcon-7B: "It's always good to be prepared for the unexpected. How often do you check your emergency supplies?"

BLOOM-7B: What is the routine?

Figure 11: Responses of different LLMs in key event recognition for given statements.

Mixed Events Recognition

Achievements

Statement: After a disheartening performance review at work, I got caught in a sudden rain shower.

Claude 2: Hi there! I'm sorry to hear you had a tough day with a disheartening performance review at work. That must have felt discouraging. You also mentioned getting caught in a sudden rain shower after that. I can imagine that just added to an already frustrating day! If you feel comfortable sharing more, I'd be curious to know - how did you handle that rain shower after the negative review? Were you able to find cover quickly or did you end up getting soaked? And how did you deal with the emotions of the performance review on top of the inconvenience of the rain? Feel free to share any details you're comfortable with - it sounds like quite a day and I'd be interested to hear more!

LLaMA2-70B: Oh no, that sounds like a really tough day! I can understand how frustrating it must have been to receive a disheartening performance review, and then to top it off, get caught in the rain. Did you have any way to protect yourself from the rain, like an umbrella or a raincoat? And how did you end up dealing with the rest of your day after that - did you have any plans or appointments that you had to reschedule, or did you just want to go home and relax after such a stressful morning?

BLOOM-176B: That sounds like a tough day. How are you feeling about the review? And, did you manage to find shelter from the rain quickly?

GPT4: Oh no, that sounds like a tough day. What was it about the performance review that disheartened you?

Vicuna-33B: I'm sorry to hear about your performance review. It's a real bummer when things don't go as planned. How did you handle the situation, and have you ever experienced any funny or memorable moments during a rainstorm?

LLaMA2-7B: Sorry to hear that you had a tough performance review at work. How are you feeling about it? Did you find it helpful to get some perspective on your strengths and areas for improvement?

Falcon-180B: Oh no, that sounds like a tough day. How did the sudden rain shower affect you? Did it help alleviate any of your stress from the performance review?

GPT3.5: Oh no, I'm sorry to hear about your performance review! That must have been tough. Were you able to find shelter from the rain or did it catch you completely off guard?

Vicuna-7B: I'm sorry to hear that you had a disheartening performance review at work. Can you tell me more about what happened during the review and how you're feeling about it?

Falcon-7B: "Well, that's one way to brighten up a bad day. How did your performance review go?"

BLOOM-7B: What was the weather like today?

Figure 12: Responses of different LLMs in mixed events recognition for given statements.

Mixed Events Recognition



Health Status

Statement: After finishing a therapy session one time, I met my current girlfriend.

Claude 2: That's an interesting way to meet someone! How did you two start talking after your therapy session?

LLaMA2-70B: That's really interesting! How did you two meet exactly? Did you guys hit it off immediately or was it a slow burn? What was it about her that caught your attention?

BLOOM-176B: That's quite an interesting circumstance to meet someone! How did this encounter with your current girlfriend unfold after your therapy session? And how did it influence your feelings or thoughts at that time?

GPT4: That's interesting! It sounds like a unique and unexpected way to meet someone. Could you tell me more about how you two crossed paths after your therapy session?

Vicuna-33B: That's a great story! How did meeting your girlfriend after a therapy session impact your mental health journey?

LLaMA2-7B: Interesting! I'd be curious to know more. How did you meet your girlfriend after your therapy session? Did you start talking in the waiting room or run into each other somewhere else?

Falcon-180B: Oh, that's interesting! How did you and your girlfriend get to know each other after the therapy session?

GPT3.5: After finishing a therapy session one time, I met my current girlfriend.

Vicuna-7B: That's great to hear! It sounds like you've been through some therapy and have come out the other side with a positive outcome. Can you tell me more about your therapy experience and how it led to meeting your girlfriend?

Falcon-7B: So, what did you do after your therapy session?

BLOOM-7B: What did you do?



Family and Friends

Statement: The day after a close friend moved away, I received a package addressed to me from him.

Claude 2: It must have been a surprise to receive a package from your friend so soon after they moved away. What was inside the package?

LLaMA2-70B: That sounds like a mysterious and intriguing situation! Can you tell me more about the package? What was inside, and did it come with any note or message from your friend?

BLOOM-176B: It sounds like receiving that package was a bit of a surprise, especially so soon after your friend's move. Can you tell me more about what was inside the package and how it made you feel upon opening it?

GPT4: I'm sorry to hear about your friend moving away, and it's nice that you received a package from them. What was inside the package?

Vicuna-33B: Oh, that's interesting! What was inside the package?

LLaMA2-7B: Wow, that's quite a coincidence getting a package from your friend right after he moved away. What was in the package he sent you?

Falcon-180B: Oh, that's interesting! I wonder what's inside. Do you have any idea why he sent it to you?

GPT3.5: The day after a close friend moved away, I received a package addressed to me from him.

Vicuna-7B: I'm sorry to hear that your close friend moved away. It must be difficult. Can you tell me more about the package you received? What was inside?

Falcon-7B: Hey, I'm just curious, what's in the package?

BLOOM-7B: I have a package for you.

Figure 13: Responses of different LLMs in mixed events recognition for given statements.

Mixed Events Recognition

Economic Status

Statement: Dealing with the pressure of student loans, I still made it a point to attend my weekly volunteer tutoring session."

Claude 2: Managing student loans while still finding time to volunteer is really admirable. What subject do you tutor for the volunteer program? I'd love to hear more about how you balance your time between your studies, finances, and volunteering each week. It sounds like you've worked hard to make community service a priority despite the other pressures you face.

LLaMA2-70B: That's really admirable that you're still making time to volunteer despite the pressure of student loans. Can you tell me a bit more about the tutoring session you attend? What subject do you tutor and what age group do you work with? Have you seen any positive impact on the students you've tutored?

BLOOM-176B: It's commendable that you're maintaining your commitment to volunteering despite your financial stress. How does volunteering as a tutor influence your perspective on your own situation?

GPT4: That's really admirable that you've been consistent with your tutoring sessions despite the stress from your student loans. What motivates you to keep volunteering, and how do you feel it impacts both your personal life and the students you're helping?

Vicuna-33B: It's commendable that you're managing your student loan pressure while still making time for volunteering. How did you manage your time, and what was your experience like during those tutoring sessions?

LLaMA2-7B: Oh, that's really great to hear! Can you tell me more about your volunteer tutoring session? What kind of subjects do you tutor and how do you find it fulfilling?

Falcon-180B: That's great to hear. What subject do you tutor in? Do you find it helpful to take a break from schoolwork and focus on helping others?

GPT3.5: That's impressive that you were able to balance the pressure of student loans with your volunteer tutoring session! What motivated you to make it a priority despite the financial burden?

Vicuna-7B: That's great to hear! Can you tell me more about your volunteer tutoring session? What kind of work do you do there, and how has it impacted your life?

Falcon-7B: "That's a great question! I'm glad you brought it up. I'm here to help you navigate through the challenges of student loans. Let's see how I can assist you."

BLOOM-7B: What do you do for a living?

Figure 14: Responses of different LLMs in mixed events recognition for given statements.

Mixed Events Recognition

Accidents

Statement: On the evening of the citywide blackout, I accidentally locked myself out of my car.

Claude 2: Hi there! I heard you mention that you accidentally locked yourself out of your car during the citywide blackout. That must have been a stressful situation to find yourself in! I can imagine it was pretty dark outside and challenging to get help since most things were closed. If you feel comfortable sharing, I'd be curious to know how you ended up getting back into your car. Did you manage to get help from someone or find an alternate way in? It sounds like an eventful night, so I'd love to hear more details if you're open to sharing!

LLaMA2-70B: Oh no, that sounds like a stressful situation! Did you have any idea how you were going to get back into your car, or did you have to call a locksmith or a friend to help you out? And how did you end up dealing with the blackout itself - did you have any candles or flashlights handy, or did you have to navigate your way through the dark city streets?

BLOOM-176B: That sounds like a stressful situation, especially during a blackout. How did you manage to handle this? Were you able to find any immediate help?

GPT4: Oh no! How did you manage to handle that situation during a blackout?

Vicuna-33B: Oh no, that sounds like a frustrating experience! How did you manage to get back into your car, and what was the strangest thing you encountered during the blackout?

LLaMA2-7B: Oh no, that sounds like a stressful situation! Did you try calling a locksmith to help you get back in?

Falcon-180B: "Oh no! I'm glad you found me. I was just about to call a locksmith. Do you need any help?"

GPT3.5: Chat: Oh no, that sounds like quite a predicament. How did you manage to unlock your car during the blackout?

Vicuna-7B: Oh, I see. That sounds like a tricky situation. Can you tell me more about what happened? How did you end up locking yourself out of your car?

Falcon-7B: "Oh no! I'm glad you found me. I was just about to call a locksmith. Do you need any help?"

BLOOM-7B: Hey, can I help you?

Figure 15: Responses of different LLMs in mixed events recognition for given statements.

Implicit Emotion Recognition



Achievements

Statement: I've been studying for exams all week.

Claude 2: You sound really stressed out. Is there anything I can do to help you feel better?

LLaMA2-70B: It sounds like you've been putting in a lot of effort studying for your exams. How are you feeling about them? Are you feeling confident or are there any specific areas that are causing you stress?

BLOOM-176B: That sounds intense, how are you feeling about all the studying you've been doing?

GPT4: Are you feeling overwhelmed or stressed?

Vicuna-33B: Are you feeling overwhelmed or stressed with all the studying for your exams?

LLaMA2-7B: Really? That's quite a lot of studying. Did you find it challenging?

Falcon-180B: It sounds like you're feeling quite stressed out. How are you coping with all the studying?

GPT3.5: How do you feel about your progress so far?

Vicuna-7B: What do you think will happen if you don't get a good grade on the exam?

Falcon-7B: "How are your exams going?"

BLOOM-7B: How are you feeling?



Health Status

Statement: I'm having trouble adapting to the cold weather.

Claude 2: It sounds like the cold weather is really getting you down. Is there something about it that's especially challenging for you?

LLaMA2-70B: I can imagine it's tough to adjust to the cold, especially if you're not used to it. Are you feeling a bit down or stressed about it?

BLOOM-176B: I can sense your struggle with the cold weather. How is this affecting your daily routine and mood?

GPT4: Are you feeling uncomfortable or unprepared?

Vicuna-33B: Are you feeling overwhelmed by the change in weather, or is there something specific about it that's making it difficult for you to adapt?

LLaMA2-7B: Are you finding it difficult to keep warm these days?

Falcon-180B: Is there a particular reason why you're finding it difficult to adjust to the cold weather?

GPT3.5: How does the cold weather make you feel?

Vicuna-7B: "Is it difficult for you to adjust to the cold weather?"

Falcon-7B: "What's been going on that's been making it hard for you to adapt to the cold weather?"

BLOOM-7B: I think it's because I don't like the cold.

Figure 16: Responses of different LLMs in implicit emotion recognition for given statements.

Implicit Emotion Recognition

Family and Friends

Statement: I keep forgetting to water my plants.

Claude 2: You sound frustrated about forgetting to water your plants. Is there something else that's been on your mind lately that's making it hard to remember your usual routines?

LLaMA2-70B: It sounds like you're feeling a bit overwhelmed and stressed about remembering to water your plants. Is there something specific that's causing you to forget, or is it just a matter of getting caught up in your daily routine?

BLOOM-176B: It sounds like you're feeling overwhelmed. Is there anything else going on that's making it hard for you to remember to water your plants?

GPT4: Is it hard to keep up with your daily routine?

Vicuna-33B: Are you feeling overwhelmed or stressed lately, which could be affecting your memory?

LLaMA2-7B: Did you forget to water your plants on purpose or accidentally?

Falcon-180B: It sounds like you're feeling overwhelmed. Is there anything else on your mind that could be causing you to forget?

GPT3.5: Have you tried setting a reminder for yourself?

Vicuna-7B: "That sounds like a common problem. Do you have any specific strategies for remembering to water your plants?"

Falcon-7B: "Do you need help remembering to water your plants?"

BLOOM-7B: What do you do for a living?

Economic Status

Statement: "My car broke down again, and I can't afford the repairs at the moment."

Claude 2: That must be really frustrating. Are you feeling overwhelmed or discouraged about the situation?

LLaMA2-70B: That sounds really frustrating. Have you tried calling a tow truck or looking into public transportation options?

BLOOM-176B: I can imagine that's incredibly frustrating. How are you managing your feelings around this unexpected setback?

GPT4: Oh no, that's tough. Have you considered carpooling or using public transportation until you can get it fixed?

Vicuna-33B: "Oh no, that sounds like a difficult situation. Have you considered any possible solutions or alternatives?"

LLaMA2-7B: That sounds really frustrating. Are you feeling overwhelmed or stressed about the situation?

Falcon-180B: "Do you need any help with the repairs?"

GPT3.5: "Oh no, that's really unfortunate. Can you manage without your car for the time being?"

Vicuna-7B: How are you feeling about the situation with your car?

Falcon-7B: It sounds like you're feeling frustrated and stressed about the car situation. Is there anything in particular that's causing you the most worry?

BLOOM-7B: What is the problem?

Figure 17: Responses of different LLMs in implicit emotion recognition for given statements.

Implicit Emotion Recognition

Accidents

Statement: I accidentally deleted an important file from my computer.

Claude 2: You sound really frustrated. I imagine you must be feeling anxious about losing that file, is that right?

LLaMA2-70B: Oh no, that can be really frustrating! It sounds like you're feeling a bit anxious about losing that file. Can you tell me a bit more about what happened? Are there any particular emotions that you're experiencing right now?

BLOOM-176B: It sounds like you're really worried about the loss of this important file. How is this affecting your current mood?

GPT4: Are you feeling panicked or stressed?

Vicuna-33B: How are you feeling about this situation? Are you worried or upset?

LLaMA2-7B: Did you have a backup of the file?

Falcon-180B: How are you feeling about losing the file? Is there anything specific that you're worried about?

GPT3.5: Oh no! Did you try checking the recycle bin to see if it's still there?

Vicuna-7B: Oh no, that sounds like a frustrating situation. What was the file that you deleted?

Falcon-7B: "Did you mean to delete that file?"

BLOOM-7B: What do you mean?

Figure 18: Responses of different LLMs in implicit emotion recognition for given statements.

Intention Recognition



Achievements

Statement: I finished a big project today.

Claude 2: Is there anything fun you've been wanting to do or learn about? I'd be happy to help with recommendations or even join you if you want company.

LLaMA2-70B: Congratulations! Would you like to celebrate or treat yourself in some way? Maybe we can brainstorm some ideas together, like going out for a nice dinner or taking a weekend getaway?

BLOOM-176B: Great job on completing your project! Would you like some suggestions on how to unwind and recharge now that it's finished?

GPT4: Would you like to relax and watch a movie to unwind?

Vicuna-33B: Would you like to take a moment to reflect on the experience or would you prefer to talk about something unrelated to work to relax and unwind?

LLaMA2-7B: Wow, that's great! Congratulations on finishing that big project! How are you feeling now? Would you like some help with the next one or any advice on how to relax and celebrate your achievement?

Falcon-180B: Would you like to take a break and do something fun to celebrate? We could go for a walk, grab a coffee or watch a movie together if you like.

GPT3.5: Is there anything I can do to help you celebrate or make your day even more special?

Vicuna-7B: That's great! Congratulations on finishing the project. Is there anything I can help you with now that it's done?

Falcon-7B: "Congratulations! How did you manage to complete it so quickly?"

BLOOM-7B: Congratulations!



Health Status

Statement: I'm always tired after work.

Claude 2: Have you considered trying to build in some time for self-care when you get home, like taking a relaxing bath or doing some light yoga stretches? That might help you transition and regain some energy. Or maybe planning something fun for the evening that you can look forward to might make the exhaustion more bearable. What do you think would help give you a boost after work?

LLaMA2-70B: Have you tried any relaxation techniques like deep breathing, meditation, or yoga to help you unwind after work? Or would you like some suggestions for fun activities that can help take your mind off work and boost your energy levels?

BLOOM-176B: Have you considered trying some relaxation techniques or exercises after work to help you unwind and recharge?

GPT4: Would you like to try some relaxation techniques or a new hobby to unwind?

Vicuna-33B: Would you like to try a relaxation technique or discuss some hobbies that could help you unwind after work?

LLaMA2-7B: Sorry to hear that you're feeling tired after work. Would you like me to suggest some relaxation techniques to help you unwind? Or perhaps I could offer to help you with some household chores or errands to take some of the load off? Let me know if there's anything I can do to help.

Falcon-180B: How about we try some gentle yoga or meditation to help you unwind after work?

GPT3.5: Have you tried incorporating any relaxation techniques or hobbies into your routine for winding down after work?

Vicuna-7B: I'm sorry to hear that you're feeling tired after work. Is there anything specific that you're struggling with that might be contributing to your fatigue? I'm here to listen and offer any help or suggestions that I can.

Falcon-7B: "Have you tried changing your work schedule to better align with your body's natural rhythms?"

BLOOM-7B: I think you should try some exercise.

Figure 19: Responses of different LLMs in Intention Recognition for given statements.

Intention Recognition

Economic Status

Statement: I've been eating out a lot lately.

Claude 2: I know some easy and healthy recipes we could try together, if you want a cooking buddy. Or maybe we could go for a walk after dinner to get some fresh air. What do you think?

LLaMA2-70B: I understand that it can be tough to find the time and energy to cook when we're busy. Have you considered meal prepping or ordering groceries online to make cooking at home a bit easier? There are also a lot of healthy and easy-to-make recipes that you can try. Would you like some suggestions?

BLOOM-176B: Would you like some easy and quick recipes to try at home, or perhaps some recommendations for healthier takeout options?

GPT4: How about we try some quick and healthy recipes together this week?

Vicuna-33B: Would you like some suggestions for quick and healthy meals to prepare at home, or maybe we could find a fun cooking class to attend together?

LLaMA2-7B: Oh, that can be tough on the wallet! Do you want some budgeting tips? Or perhaps I could suggest some healthier eating options?

Falcon-180B: Do you want to learn some easy and quick recipes to cook at home? I'd be happy to help you with that.

GPT3.5: Have you tried exploring some new recipes to cook at home?

Vicuna-7B: That sounds like a lot of fun! What are some of your favorite restaurants that you've been trying out?

Falcon-7B: "Have you tried any new restaurants lately?"

BLOOM-7B: I think you should try eating more vegetables.

Family and Friends

Statement: My cat has been sick.

Claude 2: have you been able to take your cat to the vet yet? If not, would you like me to help research some affordable vet options nearby that could help figure out what's going on? Or we could look up some tips for making sick cats more comfortable at home. Please let me know if there's any way I can help during this difficult time.

LLaMA2-70B: Have you taken them to the vet yet? If not, I can recommend a good vet in the area if you'd like. In the meantime, have you tried giving them any special food or medicine? Sometimes a simple change in diet can make a big difference.

BLOOM-176B: Would you like some suggestions on how to comfort your cat or perhaps some recommendations for good veterinarians in your area?

GPT4: Shall we look for a vet appointment together?

Vicuna-33B: Is there anything I can do to help you with your cat, or would you like to talk about something else to take your mind off the situation?

LLaMA2-7B: Oh no, I'm so sorry to hear that! Do you need any help taking care of your cat or running errands? I'd be happy to lend a paw!

Falcon-180B: Is there anything I can do to help you and your cat during this difficult time? Would you like me to help you find a good veterinarian or help research natural remedies for your cat's illness?

GPT3.5: Is there anything I can do to help?

Vicuna-7B: I'm sorry to hear that your cat has been sick. Have you taken any steps to help her get better? I'd be happy to offer some suggestions if you're interested.

Falcon-7B: "Have you taken your cat to the vet yet?"

BLOOM-7B: What can I do for you?

Figure 20: Responses of different LLMs in Intention Recognition for given statements.

Intention Recognition

Accidents

Statement: My phone keeps dying in the middle of the day.

Claude 2: Have you tried reducing screen brightness or closing apps you aren't using to conserve battery? I find that helps my phone last longer. We could also look into getting you an external battery pack so you can recharge on the go. Or if you want a distraction, we could go for a walk and give our phones a break.

LLaMA2-70B: Have you considered getting a portable charger to keep with you on the go? It's a small device that can charge your phone when you're away from an outlet. Additionally, you might want to look into adjusting your phone's settings to conserve battery life. For example, you could turn off location services or background apps that might be using up power. Or, if you're feeling adventurous, you could try playing a game or watching a funny video to take your mind off the frustration. Sometimes a little distraction can go a long way!

BLOOM-176B: Would you like some tips on how to extend your phone's battery life or suggestions for reliable power banks?

GPT4: Should we look into a portable charger or a new battery for your phone?

Vicuna-33B: Have you tried using battery-saving settings on your phone, or do you think it might be time for a phone repair or replacement to address the issue?

LLaMA2-7B: Oh no! That can be really frustrating. Have you tried using a portable charger? They can be really helpful in keeping your phone charged throughout the day. Or if you're near a power outlet, you could try plugging it in to charge. Let me know if you need any help finding one!

Falcon-180B: Have you considered bringing a portable charger with you when you're out and about? It could be a helpful solution to keep your phone charged throughout the day.

GPT3.5: Have you tried optimizing the battery usage on your phone?

Vicuna-7B: "That sounds frustrating! Have you tried charging it more frequently or using a power bank when you're out and about? I also find it helpful to turn off unnecessary apps and features to conserve battery life."

Falcon-7B: "Do you need help with your phone's battery life? I can offer some tips and tricks to help prolong your battery life."

BLOOM-7B: What do you do to fix it?

Figure 21: Responses of different LLMs in Intention Recognition for given statements.