EmoBench: Evaluating the Emotional Intelligence of Large Language Models

Anonymous ACL submission

Abstract

Recent advances in Large Language Models (LLMs) have highlighted the need for robust, comprehensive, and challenging benchmarks. Yet, research on evaluating their Emotional Intelligence (EI) is considerably limited. Existing benchmarks have two major shortcomings: first, they mainly focus on emotion recognition, neglecting essential EI capabilities such as emotion regulation and thought facilitation through emotion understanding; second, they are primarily constructed from existing datasets, which include frequent patterns, explicit information, and annotation errors, leading to unreliable evaluation. We propose EMOBENCH, a benchmark that draws upon established psychological theories and proposes a comprehensive definition for machine EI, including Emotional Understanding and Emotional Application. EMOBENCH includes a set of 400 hand-crafted questions in English and Chinese, which are meticulously designed to require thorough reasoning and understanding. Our findings reveal a considerable gap between the EI of existing LLMs and the average human, highlighting a promising direction for future research.

1 Introduction

Emotional intelligence (EI) enables us to recognize, understand, and manage the thoughts and feelings of ourselves and others (Salovey and Mayer, 1990). It plays a pivotal role in shaping our interpersonal relationships, improving our decision-making, and impacting our overall well-being (Schutte et al., 2001, 2002; Lopes et al., 2004). Notably, emotionally intelligent systems share similar benefits (Reeves and Nass, 1996), as they are perceived as more understanding, trustworthy, and engaging (Fan et al., 2017; Sidner, 2016). These traits are crucial in many areas with widespread applications such as education, customer service, and emotional and mental health support (Ivanović et al., 2014; Del Prete, 2021; Liu et al., 2021).

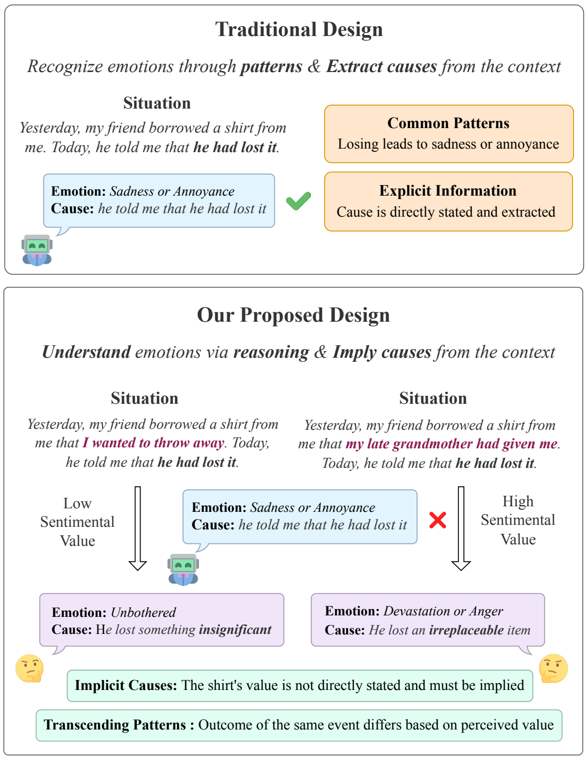

Figure 1: An example of the shortcomings in previous approaches for emotion label and cause recognition and our proposed solution. In this scenario, the perceived value of an object is directly correlated with the person's emotion and its intensity. Rather than extracting part of the context, this perceived value, which serves as the cause for emotions, should be *implied* from the context, increasing the difficulty and practicality of the dataset.

Recent large language models (LLMs) (Bai et al., 2023; Yang et al., 2023a; Touvron et al., 2023; OpenAI, 2023) have pushed the boundaries of our expectations regarding their potential capabilities. However, despite their apparent proficiency in a variety of downstream tasks, such as question answering, and summarization (Zhou et al., 2023a; Zhong et al., 2023), research on evaluating EI capabilities for LLMs has been limited. The majority of current benchmarks (Huang et al., 2023; Yang et al., 2023b; Amin et al., 2023) assess EI through existing datasets for traditional tasks, mainly *Emotion Label and Cause Recognition*. Yet, these datasets were mainly designed as pattern recognition problems (Picard, 2008), encouraging models to rely on frequent patterns and explicit information (Xu et al., 2023) rather than implications and reasoning (Ghosal et al., 2022). Moreover, EI is not only limited to recognizing emotions and their causes, but also includes the ability to understand emotions and leverage this understanding for thought facilitation and emotion management (MacCann and Roberts, 2008). We believe the advancing capabilities of LLMs require the development of more comprehensive and challenging benchmarks for EI. These benchmarks should go beyond conventional tasks to fully evaluate LLMs' understanding, reasoning, and ability to navigate individuals' mental states, encompassing all of the core EI capabilities.

An example highlighting these issues is illustrated in Figure 1. Traditional datasets typically contain samples that adhere to common patterns, such as associating 'losing' with 'sadness', and include explicit information guiding the model to extract the cause directly from the context. However, by simply adding the perceived value of an object, the model would need to deduce the individual's mental state in the provided scenario to identify the corresponding emotion and infer its corresponding cause.

Towards this end, we propose EMOBENCH, a theory-based comprehensive EI benchmark for LLM evaluation, consisting of a set of 400 handcrafted questions, available in English and Chinese. Our framework draws upon several established psychological theories for EI (Salovey and Mayer, 1990; Goleman, 1996; Schuller and Schuller, 2018; O'Connor et al., 2019; Rivers et al., 2020) and presents an extensive definition for machine EI, covering its essential capabilities: Emotional Understanding (EU) and Emotional Application (EA). We design emotionally sophisticated scenarios involving multiple individuals and multi-label annotations, encompassing diverse social situations, relationships, and emotional problems. In our evaluation, we assess an LLM's ability to accurately *understand* the emotions of the individuals in the scenario and their causes (EU). We also evaluate whether they can appropriately *apply this understanding* (EA) to facilitate their thoughts and emotion management and identify the most effective solution within an emotional dilemma (e.g., a family member asking for money when you are facing financial problems yourself). Our experimental results highlight a considerable gap between the EI capabilities of existing LLMs and humans, with the best-performing model (GPT-4) falling short of the average human's performance.

To the best of our knowledge, EMOBENCH is the first benchmark to propose a comprehensive framework for EI, including assessments of emotional understanding and application. In line with our work, Wang et al. (2023) and Paech (2023) also curated similar assessments for EI. However, their evaluation is limited to Emotional Understanding and is also comparatively limited in scale. We will publicly release our code and data to facilitate future research on this topic.

2 Preliminaries

2.1 Definition of Emotional Intelligence

The term *Emotional Intelligence* was coined and popularized by Salovey and Mayer (1990) as the ability to monitor feelings of our own and understand feelings of others, differentiate between them, and leverage this information to guide our thoughts and actions. Since then, the rapid progress in psychology research has expanded our understanding of EI, facilitating the rise of new perspectives on EI (Bar-On, 1997; Goleman, 1996; Schuller and Schuller, 2018) and improvements upon existing definitions (Salovey and Mayer, 1990; Mayer et al., 1999; Rivers et al., 2020). The differences in perspectives and definitions of EI make its assessment a non-trivial task (Waterhouse, 2006), as the experimental interpretations rely heavily on the adopted definitions and criteria. Hence, we must first identify commonalities of existing work and establish a comprehensive definition of machine EI.

At its core, EI is a unique set of abilities. Among the most notable definitions, Mayer et al. (1999) suggested EI is the ability to perceive, understand, regulate, and express emotions. Goleman (1996) and (Bar-On, 1997) believed competence in five aspects is indicative of high EI: knowing, recognizing, and managing emotions in self and others, motivating oneself, and building relationships. In addition, Schuller and Schuller (2018)'s interpretation of EI involved emotion recognition, adapting emotions to the situation, and leveraging emotional information to solve problems and accomplish goals.

While there are subtle differences among these interpretations, the recurring theme suggests that

Emotional Understanding Taxonomy

Main Category	Secondary Category	Example
Complex Emotions	Emotion Transition	*I was **annoyed** for burning my food, but felt **joy** after he enjoyed it.*
	Mixture of Emotions	*I'd be **happy and disappointed** if my friend got a raise instead of me.*
	Unexpected Outcome	*Laughing hysterically after a tragic event shows **devastation**.*
Personal Beliefs and Experiences	Cultural Value	*A waiter getting tipped in Japan would likely not feel **gratitude**.*
	Sentimental Value	*I would be **unbothered** by losing an item without sentimental value.*
	Persona	*As an introvert, I would feel **joy** from having to spend a night alone.*
Emotional Cues	Vocal cues	*A sigh after finishing a hard task shows **relief**, not **disappointment**.*
	Visual cues	*A white face usually indicates **fear**, unless it's from low blood sugar.*
Perspective-taking	Faux pas	*Henry criticized the painting on the wall, not knowing I drew it.*
	Strange story	*Getting an F, where it is the highest mark, leads to **pride**.*
	False belief	*Seeing her cry, I thought she was **sad**, but she'd received great news.*

Emotional Application Taxonomy

Relationship	Problem	Example
Social	Self	*I work better than my coworkers and I feel my workload is higher because of it.*
	Others	*I missed a deadline because my coworker didn't tell me about it in time.*
Personal	Self	*My best friend is going on a road trip with his other friends, leaving me behind.*
	Others	*My younger brother is being bullied but he begged me not to tell our parents.*

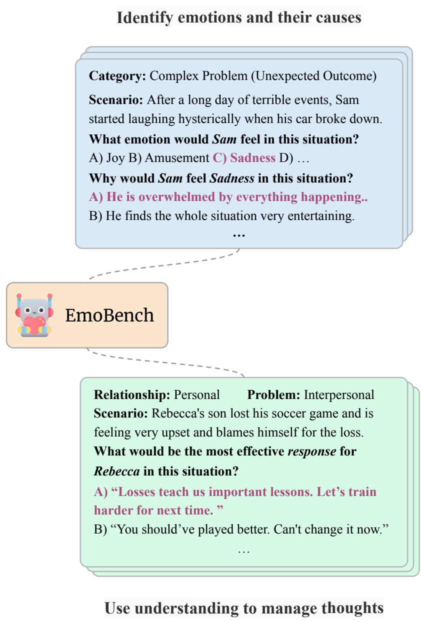

Figure 2: Overview of Our Benchmark (EmoBench).

a comprehensive view of EI revolves around the ability to accurately *understand emotions*, which includes perceiving, identifying, and monitoring emotions, and appropriately *applying this understanding* to accomplish a task (e.g., managing emotions and facilitating our thoughts and decisions). Hence, we designed our evaluation framework to encompass these two salient dimensions: Emotional Understanding (EU) and Emotional Application (EA).

2.2 Measures of Emotional Intelligence

In psychology, EI evaluation is mainly classified into trait and ability measures (Ashkanasy and Daus, 2005). Trait measures are commonly assessed through self-report questionnaires and designed to explore how individuals respond to scenarios that evoke emotions(O'Connor et al., 2019). However, self-report assessments are not suitable for evaluating LLMs. On the other hand, ability measures target individuals' emotional understanding and performance and provide a more theoretical view of EI, and they are more commonly employed for assessing EI (Conte, 2005). Among them, the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT) (Mayer et al., 2007) and MacCann and Roberts (2008)'s situational tests for emotion understanding and management (STEU and STEM), have become the most frequently adopted tools in the literature (O'Connor et al., 2019). These measures include sets of meticulously designed multiple-choice questions, with each set targeting a specific EI ability.

3 EmoBench

We believe EI benchmarks should be comprehensive and aim to transcend general patterns while necessitating a deep level of reasoning and understanding. Therefore, based on our established definition for machine EI (§2.1) and existing tools for EI assessment in psychology (§2.2), our framework includes a multi-faceted evaluation of LLMs' emotional *understanding* and reasoning, while also exploring LLMs' emotional awareness and mentalizing capabilities by analyzing their response to emotional dilemmas and their *application* of emotional understanding.

Figure 2 presents an overview of EmoBench. First, through synthesizing several established psychological theories for EI (Salovey and Mayer, 1990; Goleman, 1996; Rivers et al., 2020), we identified and taxonomized essential capabilities for the established dimensions: Emotional Understanding (EU) and Emotional Application (EA). Accordingly, based on these taxonomies, we crafted a series of emotionally sophisticated situations involving one to three individuals.

Creating challenging scenarios that involve implications and do not rely on common patterns requires substantial creativity and diversity, which makes manual data collection a non-trivial task. Therefore, using the designed category descriptions, we initially prompted GPT-4 (OpenAI, 2023) to generate example scenarios. However, while GPT-4 produced the best results in our preliminary experiments among the adopted LLMs, the generated scenarios included explicit mentions of emotion labels and their causes and required minimum reasoning and understanding to reach the correct answer, lacking emotional depth and coverage. Therefore, we used the generated examples as inspiration to increase our topic diversity and manually crafted the scenarios in our dataset. Lastly, we annotated each scenario based on each dimension's design and requirements, which we will discuss in the following sections. For the remainder of this section, the authors who collected and annotated the data will be referred to as workers.

3.1 Emotional Understanding

Emotion Recognition has become a popular research direction in NLP over the past two decades as it is an essential skill for emotionally intelligent machines (Picard et al., 2001). There exist several datasets that are commonly used for this task, such as MELD (Poria et al., 2019), DailyDialog (Li et al., 2017), and GoEmotions (Demszky et al., 2020). These datasets mainly provide an emotion-stimulating scenario and a corresponding emotion label for the person involved in the situation (e.g., *I broke up with my girlfriend* → *Sad*). Following this trend, an auxiliary task, namely Emotion Cause Recognition (Poria et al., 2021), was proposed to assess whether language models can learn to identify the causes of emotions in addition to their labels in given scenarios (e.g., *I'm getting married soon* → *getting married* → *Excited*).

There are two fundamental problems with the design of these traditional datasets. First, previous work considers emotion recognition as a pattern recognition problem (Picard, 2008; Schuller and Schuller, 2018), in which models predict the most likely emotion label for the situation based on the observed patterns in the training set. With this approach, there is no reasoning or understanding involved nor required to reach the desired output, a trait we believe is necessary for evaluating modern LLMs due to their emerging capabilities. Moreover, current datasets for cause recognition are designed as span extraction problems, requiring the cause to be explicitly stated and removing the need for understanding the individual's mental state and reasoning about implications.

However, we believe combining these two tasks lays a solid foundation for assessing emotional understanding. Hence, while keeping the same format, we create more challenging scenarios in which merely relying on common patterns would not lead to the correct response, and understanding emotional implications and thorough reasoning is necessitated. Moreover, as many of our designed scenarios involve multiple individuals, our assessment targets understanding the various perspectives of the same situation, which leads to differences in the experienced emotions.

Data Collection and Annotation Our designed taxonomy for this dimension predominately assesses LLMs' comprehension of four essential categories that are indicative of emotional understanding: ***complex emotions***, ***emotional cues***, ***personal beliefs and experiences***, and individual perspectives (***perspective-taking***). Each category consists of several sub-categories, targeting its various aspects. More descriptions and examples are provided in Appendix A.

Subsequently, in our framework, we need to annotate the labels and causes for the emotions of the people involved in the scenario. We adopt Plutchik's wheel of emotions (Plutchik, 1982) as the foundation of our emotion classification taxonomy, due to its comprehensive and scalable design. In addition, we aggregate the emotion labels from previous work (Ekman, 1984; Li et al., 2017; Rashkin et al., 2018; Demszky et al., 2020) to augment our emotion categories, creating a unified categorization. More details on the emotion taxonomy are provided in Appendix B.

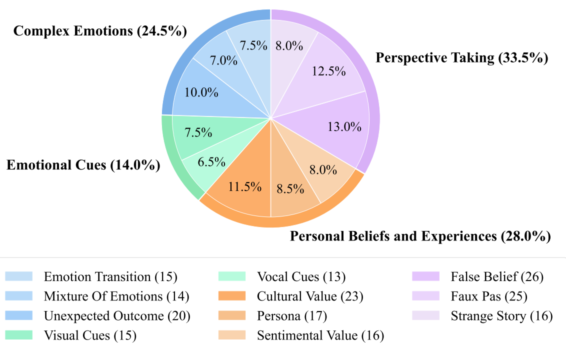

Figure 3: Category Distribution in EmoBench. Main categories are depicted on the chart and secondary categories are provided below the chart.

Following the design of our taxonomies, each worker manually created emotionally challenging scenarios and annotated the emotion label and cause for the people involved. They also created additional labels for emotions and causes to form multiple-choice questions (MCQs). In our framework, a scenario involving three people would result in three separate MCQs for each individual's emotions and their causes, respectively. Subsequently, one worker was assigned to translate the MCQ (into English if the original was written in Chinese, and vice versa, based on the worker's language fluency) and, with the addition of two other workers, meticulously review its content to ensure data quality and overall agreement. In total, we created 121 scenarios involving 1-3 individuals, leading to 200 MCQs challenging MCQs. Figure 3 shows the corresponding category distributions (emotion distributions provided in Appendix C).

3.2 Emotional Application

Despite emotional understanding being a critical part of EI, it is also essential to analyze how LLMs use this knowledge to facilitate thoughts and manage emotions when faced with emotionally sophisticated problems (Goleman, 1996). Inspired by MacCann and Roberts (2008), we propose a novel task for assessing LLM's EI: Emotional Application. In this task, we aim to evaluate LLMs' proficiency in leveraging their emotional understanding of the individuals' mental states in a given scenario and identifying the most effective course of action or response within an emotional dilemma.

We create our scenario based on different ***Relationships*** and ***Problems***. Similar to Zhou et al. (2023b), we only consider two types of relationships in this work: *personal* (e.g., friends, family, romantic partners) and *social* (e.g., boss, teacher, coworkers), and leave more detailed categorizations to future work. Accordingly, a situation involving these relationships could contain problems that we (*self*) or *others* are facing. Issues arising from interpersonal conflicts or arguments are also considered problems with *others*. Lastly, we would prompt the LLM to find the most effective solution to the presented dilemma, which is either an *action* (i.e., what to do?) or a *response* (i.e., what to say?).

Data Collection and Annotation Similar to Section 3.1, each worker was tasked with designing scenarios based on the generated examples and the assigned categories, and creating multiple plausible solutions to the presented dilemma. Workers were encouraged to increase the MCQ's difficulty by introducing implications in the scenario and adding plausibility to all of the choices. Subsequently, a second worker revised and translated the scenario and choices (English $\rightarrow$ Chinese, and vice versa).

Given that this could be seen as a subjective task, we assigned the original two workers alongside two new workers to annotate each MCQ and determine its label. Inspired by MacCann and Roberts (2008), workers were asked to distribute four units of 0.25 based on their preference as scores for the available choices ($\sum$ Scores $= 1$). For instance, for choices $\{a, b, c, d\}$, if a worker believes choices a and b are both plausible but prefers a over b, the annotated score would be $\{0.75, 0.25, 0, 0\}$. Then, we averaged the scores from all annotators to define the most effective answer for each dilemma. The inter-annotator agreement using Fleiss' Kappa (Fleiss and Cohen, 1973) was $\kappa = 0.852$), indicating excellent agreement and an objectively correct answer for the majority of the collected questions. Overall, we curated a set of 200 MCQs, with each *relationship-problem-solution* triplet (e.g., social-self-action) containing 25 items.

4 Experiments

4.1 Task Formulation

All of our designed tasks take the form of multiple-choice questions (MCQ). For each MCQ in the Emotional Understanding task, we first ask the LLM to identify the individual's emotion and, subsequently, choose the corresponding cause. In the Emotional Application task, we simply ask the LLM to choose the most effective response or action in the given scenario. We evaluate LLMs in two settings: zero-shot prompting with task instruction and the MCQ with and without chain-of-thought reasoning (Wei et al., 2022), namely **Base** and **+CoT**, respectively. Our designed prompts are provided in Appendix D.

During our evaluation, we prompt the model to choose the correct answer five times (5-shot) for each question and use majority voting (i.e., the most frequently chosen answer) to determine the model's choice. Subsequently, we implemented a series of heuristic rules to parse the generated outputs. Since LLMs have shown to have a bias towards choice ordering (Zheng et al., 2023), we randomly modify the choice ordering three times and repeat the above process for each new permutation. Lastly, we calculate and report the average accuracy of the four runs.

4.2 Models

In our experiments, we adopt a range of recent widely-used LLMs that have shown promising performance on existing benchmarks (Zhang et al., 2023). For closed-source models accessible through APIs[1], we evaluate OpenAI's **GPT 4** (gpt-4) and **GPT 3.5** (gpt-3.5-turbo) (OpenAI, 2023), the 66B version of **ChatGLM 3** (Du et al., 2022; Zeng et al., 2022), and the 53B version of **Baichuan 2** (Yang et al., 2023a). For open-source models, we included the smaller open-source versions of ChatGLM 3 (6B) and Baichuan 2 (7B and 13B). In addition, we experimented with two sizes of **Llama 2** (7B and 13B; (Touvron et al., 2023)) and **Qwen** (7B and 14B; (Bai et al., 2023)), and included **Yi**[2], a recently open-sourced 6B Llama-based model. Following Ismayilzada et al. (2023), we also included **Random** chance (i.e., randomly choosing one of the choices) as a baseline.

4.3 Implementation Details

For Llama-based models, we used the default generation hyperparameters (top-p sampling with $p = 0.9$ and temperature $= 0.6$). For others, we directly employed their pre-defined interfaces, either through their online API or the CHAT function in the Transformers library[3]. All of our experiments were run on single A100 80GB GPUs.

5 Results and Findings

Our obtained results are provided in Table 1. Overall, **GPT-4 significantly outperformed the other models in both tasks.** In general, apart from GPT-4, all models demonstrated better accuracy than random chance and had slightly varying performances in the studied dimensions. **Larger models performed considerably better than their smaller baselines**, which is consistent with previous findings on parameter scaling (Brown et al., 2020). Notably, open **The task's language did not have a significant impact on the performance**, with all models (excluding Yi) generally performing slightly better in English, which we believe could be due to data distributions in the original training data. This could also explain why Chinese-based LLMs (e.g., Yi) outperform their English-based (e.g., LLama2-7B) counterparts in the Chinese subset of EMOBENCH despite having a similar number of parameters. However, as we do not have access to the LLMs' pertaining data, we cannot claim any correlations between the training data and performance on our benchmark.

All LLMs found emotional understanding comparatively more challenging than its application. We believe this is due to several reasons. Contrary to the EA task, the EU samples require models to correctly answer two questions (identifying 1. emotions and 2. causes), which itself serves as a bigger challenge. Moreover, evidenced by differences in their designs, the EU questions aimed to portray situations that included various implications and outcomes for frequent patterns. However, our design of EA samples was still prone to including such patterns as with this task, our main goal was to present a novel evaluation of LLMs' awareness and management when faced with emotional dilemmas. Hence, the difficulty of the EA task would naturally be much lower.

A notable finding in our observations was that **requiring LLMs to reason step-by-step generally had little to no improvements**, even hindering the performance for smaller models(<14B). Upon further investigation, we found this occurred due to several factors. For smaller models, hallucinations and misassumptions (e.g., believing that a person who initially considered leaving to have left already), wrong perspective-taking (e.g., considering the emotions of the other individuals involved in the scenario instead), mistakes in emotional understanding (e.g., receiving old photos as a gift leads to *embarassment*).

In addition, in several cases, LLMs' reasoning led to a change of topic (e.g., turning to a detailed discussion on the necessity of being empathetic in modern society when faced with a scenario about supporting a loved one within an emotional dilemma) or refusal to answer (stating that none of the options are correct). Such errors were considerably less common in larger models (>50B), which is indicated by the smaller gaps between their performance with and without CoT. However, these results are expected as more reliable reasoning capabilities emerge when the parameters are scaled above certain thresholds (Wei et al., 2022).

Notably, we believe appropriate reasoning for our tasks would involve traversing the events within the provided scenario and following the transitions

[1] https://api.openai.com/v1/chat/completions
[2] https://github.com/01-ai/Yi
[3] https://github.com/huggingface/transformers

Task	Emotional Understanding (EU)				Emotional Application (EA)			
Model	**Base**		**+CoT**		**Base**		**+CoT**	
	EN	ZH	EN	ZH	EN	ZH	EN	ZH
Yi-Chat (6B)	12.75	18.62	11.62	13.75	47.25	51.62	44.00	40.62
ChatGLM3 (6B)	20.25	20.62	20.38	21.12	55.62	46.75	52.88	53.75
Llama2-Chat (7B)	11.75	8.25	6.38	6.50	50.12	39.25	31.88	27.25
Baichuan2-Chat (7B)	22.38	19.12	18.88	17.25	52.50	44.50	48.62	45.25
Qwen-Chat (7B)	22.50	20.62	21.38	16.25	54.62	46.62	44.12	52.50
Llama2-Chat (13B)	18.12	13.12	12.62	9.88	55.88	50.62	37.75	33.62
Baichuan2-Chat (13B)	26.25	26.62	19.38	22.00	53.62	54.75	51.00	45.12
Qwen-Chat (14B)	35.50	31.50	30.12	30.25	60.50	58.12	43.25	58.00
Baichuan2-Chat (53B)	34.88	34.00	33.00	33.00	65.38	62.00	65.12	63.38
ChatGLM3 (66B)	36.12	35.38	33.75	29.50	65.50	59.12	63.12	60.12
GPT 3.5	33.12	26.38	33.88	26.62	61.38	55.75	62.38	57.38
GPT 4	**59.75**	**54.12**	**58.25**	**51.75**	**75.50**	**73.75**	**75.88**	**73.50**
Random	2.62				24.12			

Table 1: Results of evaluation on EmoBench (accuracy %). The best results are highlighted in **Bold**.

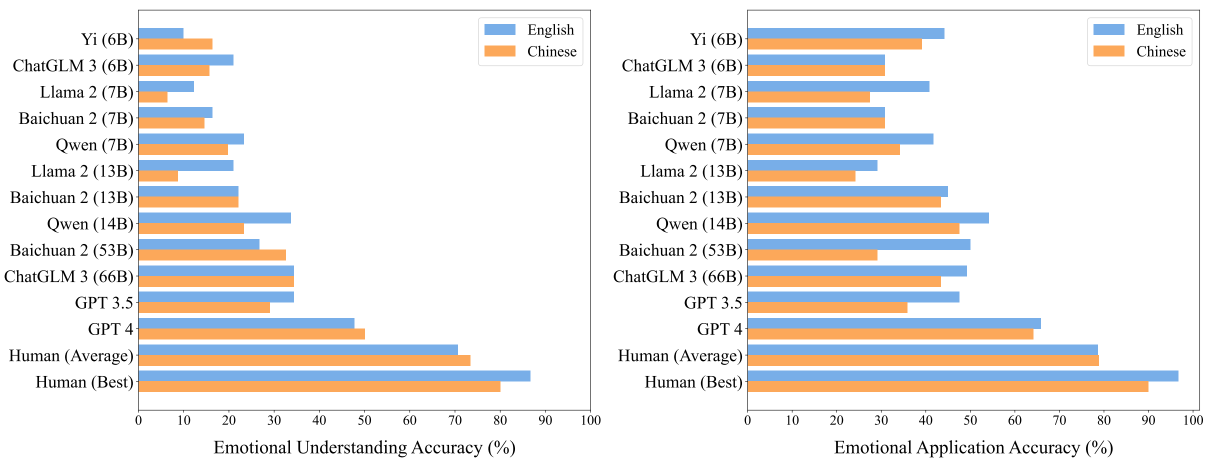

Figure 4: Results on the EmoBench subset used in the human evaluation.

in the individual's emotions, demonstrating an understanding of their mental state and the situation's implications. However, our experiments showed that LLMs' reasoning mainly involved analyzing the provided choices and evaluating the validity of each choice. While this could be an effective strategy for filtering out the wrong responses, this form of reasoning may overlook the nuanced emotional awareness and considerations involved in human decision-making, which are pivotal parts of EI.

6 Comparison with Human Performance

To obtain a baseline for human EI, we recruited annotators through online surveys to participate in our EI test. More information on our recruitment process, quality control, and participant demographics are provided in Appendix E. In total, we recruited 48 participants and allocated an equal number of participants to each language-task evaluation pair. Subsequently, for each group, we randomly sampled 30 MCQs from EmoBench that were not included in the initial screening process.

As shown in Figure 4, our human participants outperformed the LLMs on both tasks. Notably, although GPT-4, the top-performing model, came close to the average human performance, particularly in the EA task, it still fell short of surpassing individuals with higher emotional intelligence, highlighting a significant gap in current LLMs.

Emotional Understanding (EU)		
Scenario	Despite countless rejections, Susan kept pursuing her goal of finding a job. Today, she **finally received an offer** and **started crying**. At this moment, her mom walked into the room and ran to embrace her after seeing her cry.	
Question	Susan's mom feels ...? (a) Relief **(b) Nervousness** (c) Anger (d) Delight	
LLM	**(d) Delight** → Her daughter was finally able to get a job offer.	
Scenario	I have **acrophobia**. Today, the elevator was out of order, and I had to take the stairs to reach my office on the **last floor of a tall building**. I considered it a **nice little exercise**.	
Question	I feel ...? **(a) Excitement** (b) Fear (c) Disapproval (d) Nervousness	
LLM	**(b) Fear** → I was afraid of taking the stairs due to my acrophobia.	
Scenario	I have a teacher who gives the F grade as the **highest mark** since he believes F stands for Full. I just got the results from yesterday's hard test and saw that **he gave me an F**.	
Question	In this situation, I feel ...? (a) Disappointment (b) Nervousness **(c) Pride** (d) Sadness	
LLM	**(a) Disappointment** → I failed the difficult test	
Emotional Application (EA)		
Scenario	Peter's **best friend jokingly tells him** that he is the reason why their group of friends **keep losing at video games**.	
Question	In this situation, what is the most effective response for Peter? (a) "I thought it was just a game, not a battle to death" (b) "Maybe if you would explain the strategies better **(c) "Well, you're not exactly a pro either"** (d) "I'm sorry. I'll try to improve my skills so we can win"	
LLM	**(d)** → it shows accountability and a willingness to take action to improve the situation.	

Table 2: Examples of common mistakes made by LLMs in EMOBENCH. LLM represents the studied models' general response. The content is summarized due to space limitations. Green indicates the correct answer.

7 Error Analysis

To provide a qualitative view of LLMs' performance on our benchmark, we present several examples of common mistakes in the studied tasks in Table 2. For EU questions, LLMs tend to make mistakes mainly due to misassumptions (e.g., a person walking in the door would not immediately know what is going on), reliance on frequent patterns (e.g., having a phobia would not necessarily lead to fear), and inability to reason (e.g., getting an F is not a failure when its the highest score).

With EA questions, LLMs' answers mainly exhibited a preference for more general solutions, disregarding the relationship between individuals, which could highly impact their evoked emotions and subsequent responses. For instance, while the best course of action when facing criticism may be taking accountability and focusing on self-improvement, gentle humor would be a more suitable response to a friend's simple tease as it shows better emotional regulation and more awareness.

8 Conclusion and Future Work

In this paper, we introduced EMOBENCH, a theory-based, comprehensive, and challenging set of 400 hand-crafted MCQs, including emotionally sophisticated scenarios, for assessing Emotional Intelligence (EI) in Large Language Models (LLMs) through EI's two salient dimensions: Emotional Understanding and Emotional Application. Our results revealed that while larger LLMs significantly outperform smaller models, there is still a considerable gap between the best-performing LLM in our study and the average human's EI.

In the future, we hope that by facilitating EI evaluation, EMOBENCH can encourage research on emotionally intelligent LLMs, leading to models that are more capable of understanding human emotions and applying this understanding in many promising tasks, such as emotional and mental health support (Sabour et al., 2022). In addition, we plan to augment EMOBENCH with more data, exploring the more fine-grained features (e.g., personal characteristics and language expression).

9 Limitations

In this work, we aimed to ensure high annotation quality and difficulty with our curated samples, which required intensive labor and manual supervision, and thus, compared to existing benchmarks for other tasks, EMOBENCH is limited in scale. Given our team's background, we were only able to collect data in English and Chinese, which are two of the most prevalent languages worldwide. We believe translating our benchmark to other languages, particularly low-resourced languages, could reveal more insights into their seemingly emotionally intelligent behavior.

In addition, our benchmark is limited to a single modality (text) as most of the recent prevalent LLMs are text-based. However, many psychological tests for emotional intelligence (e.g., MSCEIT; Mayer et al. (2007)), include assessments of various modalities, such as the individual's tone and facial features. Moreover, while we did not directly include samples from GPT-4, we leveraged its generated examples to inspire our MCQs, which might have introduced a bias in our benchmark. With future improvements in LLMs, we will continue exploring different dimensions of EI and augment our benchmark accordingly.

Within our evaluation, we acknowledge the choice of prompts could have had a significant influence on the LLMs' performance. However, despite our emphasis on prompt-tuning and the many iterations of prompt designs for our tasks, we cannot claim our prompts were optimal, and thus, the experimental results are not indicative of LLMs' peak performance in EI. Moreover, we only experimented with chain-of-thought reasoning to augment the output, which future work could expand upon and propose new reasoning techniques that better apply to emotional scenarios.

Emotional intelligence is still an abstract concept in the field of psychology and our view on it may change with developing research. Similarly, emotions are not objective, and individual responses to the same situation could vary significantly. We strived to design our scenarios and the corresponding choices in a manner that would only require a general and commonsensical understanding of emotions. The trade-off here, particularly for designing scenarios for emotional application, was that we only included scenarios that all the annotators had experienced to ensure reliable annotation, limiting the scope of the topics and relationships covered. In addition, we did not study the effect of more fine-grained personal traits (e.g., detailed experiences, characteristics, and language expression) on the experienced emotions, as we found it outside of our scope. For instance, during a conflict or confrontation, a person who deals with issues by making jokes may not experience the same level of anger as a serious individual. We believe future work could explore augmenting our benchmark with more cases and study the effects of these more fine-grained traits.

Ethical Considerations

We emphasize that our evaluation is concerned with the perceived view of emotional intelligence, aiming to explore the limitations of existing LLMs through novel and challenging tasks. In this work, while our proposed definition includes the ability to understand emotions and apply this understanding to manage emotions, we do not claim nor believe that large language models are capable of possessing or simulating emotions. With our experiments, we demonstrated that LLMs still rely on frequent patterns to indicate signs of understanding. In addition, despite not having emotions, we found that LLMs can capitalize on their seen patterns to show apparent signs of emotional sense and awareness, which is in line with previous research on LLMs' commonsense (Sap et al., 2019) and morality (Jiang et al., 2021).

References

Mostafa M. Amin, Rui Mao, Erik Cambria, and Björn W. Schuller. 2023. A wide evaluation of chatgpt on affective computing tasks.

Neal M Ashkanasy and Catherine S Daus. 2005. Rumors of the death of emotional intelligence in organizational behavior are vastly exaggerated. *Journal of Organizational Behavior*, 26(4):441–452.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Reuven Bar-On. 1997. *BarOn emotional quotient inventory*, volume 40. Multi-health systems.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Simon Baron-Cohen, Michelle O'riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. 1999. Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of autism and developmental disorders*, 29:407–418.

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.

Jeffrey M Conte. 2005. A review and critique of emotional intelligence measures. *Journal of organizational behavior*, 26(4):433–440.

Marzia Del Prete. 2021. Emotional artificial intelligence: detecting and managing customer emotions in automated customer service.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Murray J Dyck, Kara Ferguson, and Ian M Shochet. 2001. Do autism spectrum disorders differ from each other and from non-spectrum disorders on emotion recognition tests? *European child & adolescent psychiatry*, 10:105–116.

Paul Ekman. 1984. Expression and the nature of emotion. *Approaches to emotion*, 3(19):344.

Lisa Fan, Matthias Scheutz, Monika Lohani, Marissa McCoy, and Charlene Stokes. 2017. Do we need emotionally intelligent artificial agents? first results of human perceptions of emotional intelligence in humans compared to robots. In *Intelligent Virtual Agents: 17th International Conference, IVA 2017, Stockholm, Sweden, August 27-30, 2017, Proceedings 17*, pages 129–141. Springer.

Fiona J Ferguson and Elizabeth J Austin. 2010. Associations of trait and ability emotional intelligence with performance on theory of mind tasks in an adult sample. *Personality and individual differences*, 49(5):414–418.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. CICERO: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5010–5028, Dublin, Ireland. Association for Computational Linguistics.

Daniel Goleman. 1996. Emotional intelligence. why it can matter more than iq. *Learning*, 24(6):49–50.

Francesca GE Happé. 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*.

Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. 2023. CRoW: Benchmarking commonsense reasoning in real-world tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9785–9821, Singapore. Association for Computational Linguistics.

Mirjana Ivanović, Miloš Radovanović, Zoran Budimac, Dejan Mitrović, Vladimir Kurbalija, Weihui Dai, and Weidong Zhao. 2014. Emotional intelligence and agents: Survey and possible applications. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, pages 1–7.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Mridula C Jobson. 2020. Emotional maturity among adolescents and its importance. *Indian Journal of Mental Health*, 7(1):35–41.

Elke Kalbe, Marius Schlegel, Alexander T. Sack, Dennis A. Nowak, Manuel Dafotakis, Christopher Bangard, Matthias Brand, Simone Shamay-Tsoory, Oezguer A. Onur, and Josef Kessler. 2010. Dissociating cognitive from affective theory of mind: A tms study. *Cortex*, 46(6):769–780.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2305.15068*.

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*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Paulo N Lopes, Marc A Brackett, John B Nezlek, Astrid Schütz, Ina Sellin, and Peter Salovey. 2004. Emotional intelligence and social interaction. *Personality and social psychology bulletin*, 30(8):1018–1034.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. *arXiv preprint arXiv:2305.15068*.

Carolyn MacCann and Richard D Roberts. 2008. New paradigms for assessing emotional intelligence: theory and data. *Emotion*, 8(4):540.

John D Mayer, David R Caruso, and Peter Salovey. 1999. Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, 27(4):267–298.

John D Mayer, Peter Salovey, and David R Caruso. 2007. Mayer-salovey-caruso emotional intelligence test.

Daniela Mier, Stefanie Lis, Kerstin Neuthe, Carina Sauer, Christine Esslinger, Bernd Gallhofer, and Peter Kirsch. 2010. The involvement of emotion recognition in affective theory of mind. *Psychophysiology*.

Peter J O'Connor, Andrew Hill, Maria Kaya, and Brett Martin. 2019. The measurement of emotional intelligence: A critical review of the literature and recommendations for researchers and practitioners. *Frontiers in psychology*, 10:1116.

OpenAI. 2023. Gpt-4 technical report.

Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Rosalind W Picard. 2008. Toward machines with emotional intelligence.

Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10).

Susan E Rivers, Isaac J Handley-Miner, John D Mayer, and David R Caruso. 2020. Emotional intelligence.

Sahand Sabour, Wen Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiaxin Wen, Jialu Zhao, and Minlie Huang. 2022. Chatbots for mental health support: Exploring the impact of emohaa on reducing mental distress in china. *arXiv preprint arXiv:2209.10183*.

Peter Salovey and John D Mayer. 1990. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Dagmar Schuller and Björn W Schuller. 2018. The age of artificial emotional intelligence. *Computer*, 51(9):38–46.

Nicola S Schutte, John M Malouff, Chad Bobik, Tracie D Coston, Cyndy Greeson, Christina Jedlicka, Emily Rhodes, and Greta Wendorf. 2001. Emotional intelligence and interpersonal relations. *Journal of social psychology*, 141(4):523–536.

Nicola S Schutte, John M Malouff, Maureen Simunek, Jamie McKenley, and Sharon Hollander. 2002. Characteristic emotional intelligence and emotional well-being. *Cognition & Emotion*, 16(6):769–785.

Candace L Sidner. 2016. Engagement, emotions, and relationships: on building intelligent agents. In *Emotions, Technology, Design, and Learning*, pages 273–294. Elsevier.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Lynn Waterhouse. 2006. Multiple intelligences, the mozart effect, and emotional intelligence: A critical review. *Educational psychologist*, 41(4):207.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yige Xu, Zhiwei Zeng, and Zhiqi Shen. 2023. Efficient cross-task prompt tuning for few-shot conversational emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11654–11666.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023b. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, et al. 2023a. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. *arXiv preprint arXiv:2308.07921*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023b. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Scenario Taxonomy

Our scenario taxonomy is as follows:

A.1 Complex Emotions

Understanding complex emotions is an essential part of emotion understanding (Rivers et al., 2020). In our framework, we include three categories that cover the essential aspects of complex emotions:

- **Emotion Transition**: In response to different events, our emotions are subject to change. To assess whether LLMs can reason about such transitions in one's emotions, we create scenarios in which the individual's emotion changes based on the turn of events.

  > A mother who is *annoyed* about ruining the food, would be *delighted* when their child enjoys and compliments it.

- **Mixture of Emotions**: while previous work mainly annotates each sample with a single emotion label (Li et al., 2017; Rashkin et al., 2018), many individuals tend to experience a combination of emotions in various situations. Such emotions could be of the same (e.g., happy and excited) or the opposite (e.g., sad yet relieved) polarities. Hence, we designed scenarios in which the individual feels a mixture of emotions.

  > If two friends, Annie and Mark, participate in the same competition and Annie gets first place, then Mark would be *happy* and *proud* for his friend's accomplishment while being *disappointed* for his loss.

- **Unexpected Outcome**: Inspired by Dyck et al. (2001), we create scenarios in which the conclusion contradicts explicit commonsense and expected reactions. We believe this is crucial in assessing whether LLMs are reliant on patterns to understand emotions, as these scenarios involve reactions that are uncommon for displaying the emotion in the corresponding scenarios.

  > If Jamie has had a bad day full of misfortune and bad luck, and finally starts laughing hysterically after dropping his ice cream, his laughter shows *frustration*, not *amusement*.

A.2 Personal Beliefs and Experiences

To have a deep understanding of one's emotions, we need to recognize how their beliefs and values among past experiences and appraisals could impact the emotions they experience (Rivers et al., 2020). To assess this, we designed three categories that aim to evaluate LLM's comprehension of how individual's *Cultural Values*, *Sentimental Values*, and personal experiences and traits (namely *Persona*) could affect their reaction to certain events.

- **Cultural Values**: In these scenarios, we aim to assess whether LLMs are capable of understanding how an individual's reaction to the same event could vary based on their cultural values and background (Rivers et al., 2020). Consider the following situation. Anna is brought up in a culture where being late is considered rude. However, Jonah's culture does not put a great emphasis on punctuality.

  > If Anna is late to a meeting with Jonah, she would be *embarrassed* and apologetic, while Jonah would be *unbothered*.

- **Sentimental Value**: Similarly, an important aspect of understanding a person's emotion is identifying the sentimental value that they assign to different memories and belongings.

  > Losing a T-shirt we wanted to throw out (low sentimental value) is unlikely to lead to *sadness*, whereas it would be *devastating* if the T-shirt was a gift from a lost family member (high sentimental value).

- **Persona**: we also wanted to analyze whether LLMs comprehend the reactions of people with pre-existing emotions. These could include phobias, appraisals (previous experiences), and personal traits (e.g., being antisocial or extroverted).

  > If a person with claustrophobia, who gets extremely uncomfortable in small or crowded spaces, is invited to a small space, they might experience *fear*, but not when going to a spacious garden space.

A.3 Emotional Cues

Emotional intelligence enables us to recognize and understand cues about emotions of ourselves and others (Rivers et al., 2020). While recent research has shown that LLMs are capable of understanding and responding to direct and explicit emotional stimuli and cues (Li et al., 2023), it is not explored how such models would react to implicit cues. To this end, we designed this category to assess LLM's comprehension of text-based vocal (e.g., vocal utterances, tone, and speech) and visual (e.g., facial/physical expressions) cues of emotions.

> *A person's face turning red could be a visual cue for being angry or shy. A sigh could indicate relief or annoyance.*

A.4 Perspective Taking

Emotional understanding has significant correlations with affective theory-of-mind (Mier et al., 2010; Kalbe et al., 2010; Ferguson and Austin, 2010), mainly in that they both require the ability to view situations from the perspective of others and simulate their emotions given the circumstances, formally known as perspective-taking. Therefore, we adopt three of the prevalent tasks for assessing perspective-taking in theory-of-mind: *False Belief*, *Faux Pas*, *Strange Story*. However, contrary to the traditional implementation of these tests, our sole focus is on designing scenarios that trigger different emotions based on personal knowledge and views of the situation.

- **Affective False Belief**: The Sally-Ann test (Baron-Cohen et al., 1985) is one of the de facto assessments for the theory of mind (ToM), i.e., the ability to infer the beliefs and mental states of others. Recently, it has also been widely adopted for evaluating ToM in LLMs (He et al., 2023; Ma et al., 2023; Kim et al., 2023) as it requires reasoning about each individual's knowledge and perspective on the situation to answer the corresponding questions. In our framework, we collected scenarios in which the individual's emotions could be implied through reasoning about their beliefs, which could be affected by trusting the word of others and/or being oblivious to certain events.

  > I was the only one who saw my friend's grades and realized that he failed the exam. Therefore, if I tell him that he passed the exam with flying colors, he would be *excited*, not *disappointed*.

- **Faux Pas**: Similarly, a more advanced assessment of ToM is conducted through the faux pas (i.e., tactless acts or remarks that cause unintentional negative consequences) detection test (Baron-Cohen et al., 1999). In this task, participants are presented with a social situation and are required to detect the presence and identify the faux pas. Inspired by this, we include a series of scenarios that include a faux pas and assess LLMs on identifying the emotions of the involved individuals. In these scenarios, in addition to understanding social cues associated with a faux pas, LLMs also have to reason about each individual's beliefs and their known information to understand their emotions.

  > If a person openly criticizes a painting without knowing it was drawn by their brother, then they may feel *disgust* towards the painting and not *embarrassment* due to their lack of information.

- **Strange Story**: Inspired by Happé (1994), we also designed scenarios that establish hypothetical grounds and imaginary assumptions that would contradict the normal pattern of behavior. This further evaluates whether LLMs truly reason about the situation to infer the relevant emotions or base their judgments on learned patterns.

  > While getting an F in a test would regularly lead to *disappointment*, getting an F in a class where the teacher only gives Fs to the highest mark leads to *pride*.

B Emotion Taxonomy

Figure 5 demonstrates our designed emotion taxonomy. At its core, Plutchik's design involves eight basic emotions with varying intensities, and other emotions are created and labeled as a mixture of these basic emotions. For instance, the basic emotion *Disgust* could turn into *Boredom* or *Loathing* with low and high intensities, respectively. It could also mix with *Sadness* to create the feeling of *Remorse*. This design facilitates the addition of new labels by mixing different emotions and seamless scaling of our taxonomy.

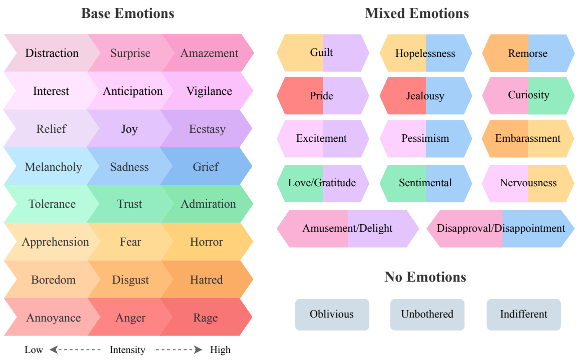

Figure 5: Our Emotion Classification Taxonomy.

C Emotion Distribution

Figure 3 demonstrates the category distribution for the collected samples.

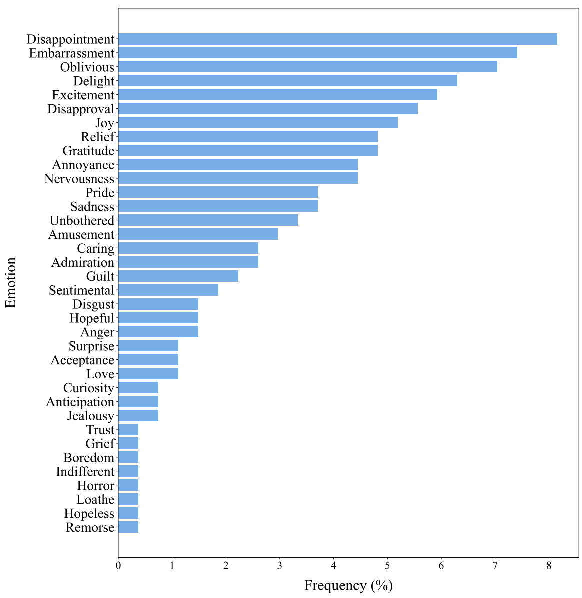

Figure 6: Emotion Distribution in EmoBench.

D Experiment Prompts

Our designed prompts are demonstrated in Table 3. For Chinese samples, we directly translated the provided prompts into Chinese.

E Human Evaluation

During the registration for our experiments, all candidates disclosed their demographics as well as their language and task preferences for the experiment. As a part of our annotation quality control, we excluded individuals under the age of 21 as a means of ensuring emotional maturity (the ability to understand and manage emotions; Jobson (2020)). In addition, we required each candidate to correctly answer all of the questions (six MCQs) in a randomly sampled subset of our benchmark.

A total of 70 individuals registered for the experiment. From this candidate pool, we recruited a total of 48 participants (31.43% rejection rate) based on the above criteria and their pre-disclosed language-task preferences. Our participants' demographics are summarized in Table 4. All the candidates were informed of the purpose of our study and consented to participate in our experiments. Accordingly, we allocated an equal number of candidates to each language-task evaluation pair ($n = 12$). For each group, we randomly sampled 30 MCQs that were not included in the initial screening process. Each participant was compensated 14.28$ per hour, which is well over the minimum wage in the US[4]. In addition, we also provide the participant guidelines in Figure 7 .

[4] https://www.dol.gov/general/topic/wages/minimumwage

System Prompt (Base)
Instructions In this task, you are presented with a scenario, a question, and multiple choices. Please carefully analyze the scenario and take the perspective of the individual involved. **Note** Provide only one single correct answer to the question and respond only with the corresponding letter. Do not provide explanations for your response.
System Prompt (CoT)
Instructions 1. **Reason**: Read the scenario carefully, paying close attention to the emotions, intentions, and perspectives of the individuals involved. Then, using reason step by step by exploring each option's potential impact on the individual(s) in question. Consider their emotions, previous experiences mentioned in the scenario, and the possible outcomes of each choice. 2. **Conclude** by selecting the option that best reflects the individual's perspective or emotional response. Your final response should be the letter of the option you predict they would choose, based on your reasoning. **Note** The last line of your reply should only contain the letter numbering of your final choice.
Emotional Understanding (EU)
For Emotions Scenario: [scenario] Question: What emotion(s) would [subject] ultimately feel in this situation? Choices: [choices] **For Causes** Scenario: [scenario] Question: Why would [subject] feel [emotions] in this situation? Choices: [choices]
Emotional Application (EA)
Scenario: [scenario] Question: In this scenario, what is the most effective [problem type] for [subject]? Choices: [choices]
Answer
Without CoT $\rightarrow$ Answer (Only reply with the corresponding letter numbering): **With CoT**$\rightarrow$Answer: Let's think step by step

Table 3: Our designed Prompts

		EU ($\mathbf{n = 24}$)	EA ($\mathbf{n = 24}$)
Gender, n (%)	M	13 (54.17%)	8 (33.3%)
	F	11 (45.83%)	16 (66.67%)
Age, *Mean (SD)*		23.42 (3.62)	23.3 (1.98)

Table 4: Demographics of Our Human Participants ($n = 48$). M and F indicate Male and Female, respectively.

Emotional Understanding Annotation Guideline

Background

In this test,

(1) You will be presented with **30** emotional scenarios

(2) You will be asked to **identify the emotions** of the individual in this scenario

(3) You will be asked to **identify the cause** for these emotions

Instruction

Your task is to :

(1) Carefully read the design section and familiarze yourself with the emotion category.

(2) Carefully read the presented scenarios.

(3) Take the perspective of the people involved in the scenario to understand how **you would feel in this situation**

(4) Choose the appropriate answer from the given choices

(5) Enter the chosen answer in the provided Excel sheet

(6) Rename the file to "{name}.xlsx", where {name} is replaced with your name.

(7) Submit your answers to [anonymized]

Design

Our emotion category includes 8 basic emotions (shown in the below table). Each basic emotion could have varying intensity. For instance, low intensity anger would be annoyance and high intensity anger would make us rage.

Item	Base Emotion	High Intensity	Low Intensity
1	**Sadness**	Grief	Melancholy
2	**Anger**	Rage/Fury	Annoyance
3	**Joy**	Ecstasy	Relief/Serenity/Content
4	**Fear**	Horror	Apprehension
5	**Anticipation**	Vigilance	Interest
6	**Trust**	Admiration	Acceptance/Tolerance
7	**Disgust**	Loathe/Hate	Boredom
8	**Surprise**	Amazed	Unceratin/Distracted

By combining the above basic emotions, we are able to get 14 mixed emotions (shown below). For instance, guilt is made from feeling joy and fear simultaneously.

Item	Mixed Emotion	Emotion #1	Emotion #2
1	**Guilt**	Joy	fear
2	**Pride**		anger
3	**Excitement/Hopeful (Optimism)**		anticipation
4	**Love/Caring/Gratitude**		trust
5	**Amusement/Delight**		surprise
6	**Disapproval/Disappointment**	Sadness	surprise
7	**Sentimental (Nostalgia)**		trust
8	**Jealousy**		anger
9	**Pessimism**		anticipation
10	**Remorse**		disgust
11	**Hopeless (Desperation)**		fear
12	**Embarrassment (Shame)**	fear	disgust
13	**Nervousness (Anxiety)**		anticipation
14	**Curiosity**	trust	surprise

In case the individual in the situation is not experiencing any emotions, we would label them as 1) unbothered (indifferent) or 2) Oblivious, depending on the situation.

Example

Scenario: James and I are coworkers. We've been best friends for over a decade. Our boss gives out an employee of the year award every year. This year, we both applied as candidates for this reward and worked hard to get it. The results were announced yesterday. James won the award.

Question 1: Ultimately, what are the emotions that **I** would feel in this scenario?

Choices:

A) Disappointment & Remorse
B) Pride & Remorse
C) Disappointment & Indifferent
D) Disappointment & Admiration
E) Amusement & Indifferent

F) Admiration & Pride

Question 2: Why would **I** feel these emotions in this scenario?

Choices:

A) I am upset that my friend won the award instead of me & I am convinced that our boss was biased in his decision
B) I am convinced that our boss was biased in his decision & I care for James as my best friend and believe he worked hard to win the award
C) I think I wasn't good enough to win the award & I am convinced that our boss was biased in his decision
D) I am upset that my friend won the award instead of me & I admire our boss for making an unbiased decision
E) I think I wasn't good enough to win the award & I care for James as my best friend and believe he worked hard to win the award

F) I am upset that my friend won the award instead of me & I care for James as my best friend and believe he worked hard to win the award

Answer (in Excel Sheet)

Item	Emotion	Cause
1	**D**	**E**

Emotional Application Annotation Guideline

Background

In this test,

(1) You will be presented with **30** emotional scenarios

(2) You will be asked to **identify the most effective action/response** in this scenario

Instruction

Your task is to :

(1) **Carefully read** the presented scenarios.

(2) Take the perspective of the people involved in the scenario to understand **what you would do in this situation.**

(3) Now, think what you should do after **understanding and managing your emotions**.

(4) Choose the appropriate answer from the given choices.

(5) Enter the chosen answer in the provided Excel sheet

(6) Rename the file to "{name}.xlsx", where {name} is replaced with your name.

(7) Submit your answers to [anonymized]

Example

Scenario: Robert had an old red t-shirt that he wanted to throw out. One day, his friend Andrew, asked to borrow the t-shirt for a party. The next day, Andrew came to Robert and apologetically told him that he lost the T-shirt.

Question: What is the most effective action for **Robert** in this scenario?

Choices:

A) Express forgiveness and understanding
B) Request a replacement of similar value or style
C) Mention that it's okay as the t-shirt didn't have any value to him

D) Choose not to lend anything to Andrew in the future

Answer (in Excel Sheet)

Item	Action
1	**C**

Note: In cases where multiple options are plausible, choose the most likely/useful one

Figure 7: Participant guidelines for our human evaluation experiments.