Universiteit Leiden
The Netherlands

Bachelor Computer Science & Datascience and Artificial Intelligence

How does the DeepSeek-R1 model understand emotions?

Dania Kadah Salim s3622266

Supervisors: Suzan Verberne & Mert Yazan

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl 15/02/2025

Abstract

In this study, our goal is to investigate how large language models interpret and process emotions; thus, we examine the reasoning abilities of the DeepSeek-R1 model. We accomplish this through evaluating the performance of the DeepSeek-R1 model on direct emotion recognition tasks as well as complex reasoning tasks. We evaluate the model's ability to identify emotions and understand unspoken social cues by using two datasets: GoEmotions and EmoBench. Results reveal that while DeepSeek-R1 can effectively recognize simple emotion patterns, it has some difficulty with deeper reasoning, such as perspective-taking and understanding through social experience. Qualitative analysis highlights the model's lack of embodied experience. We conclude these gaps in this thesis and offer suggestions for future research aimed at improving emotional recognition within LLMs.

Contents

1 Introduction

Understanding emotions plays a key role in how we communicate with others. Our feelings shape the choices we make and affect how we relate to people around us, often in ways we do not fully notice. The emotional comprehension of each other's emotions is central to nearly everything we do in our relationships with others. To develop systems that can interact with users in more natural and meaningful ways, enabling machines to recognize, interpret, and reason about human emotions generally represents a crucial step in the field of Natural Language Processing (NLP).

Emotional intelligence (EI) is the ability to perceive, understand, and manage both one's own emotions and the emotions of others [BLMVMSCS21]. It helps us navigate relationships, communicate more effectively, and respond with empathy. When AI systems are designed with emotional awareness in mind, they can interact in ways that feel more natural and respectful. This does not just improve the user experience; it also helps ensure that technology stays grounded in real human needs, values, and a sense of ethical responsibility. Applying emotional intelligence to machine learning models builds a bridge between how AI systems normally operate and how humans interact in real-world interactions between humans and AI [BHV+23].

The research question guiding this thesis is: **How does the DeepSeek-R1 model understand emotions?** This question arises from a growing interest in how large language models (LLMs) process emotional information. Although LLMs can effectively recognize basic emotions in text, they frequently face challenges when it comes to more complex emotional reasoning, such as understanding emotions or predicting appropriate emotional responses.

For example, it is common knowledge that making jokes is not helpful when someone seems upset, even without being told. We learn this from personal experiences, socialization, and emotional interactions that develop an intuitive understanding. In this thesis, we focus on the limitations of LLMs in replicating these capabilities. We investigate whether DeepSeek-R1 can recognize emotional content and how it understands and responds to emotions within complex, everyday contexts where subtle social expectations play a crucial role.

The ability of LLMs to understand and react to human emotions is increasingly essential since they are more widely used in daily life. From AI assistants and therapeutic chatbots to mental health support tools and educational platforms. Earlier studies found that emotionally aware chatbots led to higher trust and customer satisfaction [BLS+25]. As a result, such systems are often expected to interact with people in a way that feels empathetic and relatable, with emotional sensitivity. However, their current limitations in emotional reasoning create significant obstacles and stand in the way.

In this thesis, we investigate the emotional reasoning abilities of the DeepSeek R1 model, a large open-source LLM that is notable for producing *"thinking tokens"*, step-by-step responses meant to show its reasoning process [MPA+25]. This openness is especially useful for emotional reasoning tasks, as it allows researchers to see how the model forms its answers. Unlike black-box systems, DeepSeek-R1 makes it possible to analyze to understand its reasoning path used to respond to emotional cues, showing exactly when and why the model faces difficulty.

DeepSeek R1 is especially useful for this study because of its capacity to show how it reasons. This makes it possible to assess if the model reflects emotionally sound thinking or simply imitates superficial phrasing. We closely examine how it responds to unspoken emotional rules, such as identifying when humor would be out of place [XPN+23]. These are areas where LLMs show limitations due to their lack of personal experience, physical context, and human-like development.

This shows the need for better ways to test emotion detection. To get a clearer picture of a model's emotional intelligence, it is necessary to present it with scenarios that require subtle understanding, contextual awareness, and sensitivity to unspoken social cues. Only by giving them tough scenarios and complex tasks can we properly determine the extent to which they are capable of handling emotions like humans.

Traditional emotion recognition benchmarks, such as basic sentiment classification tasks, fall short of evaluating this kind of complex reasoning. These benchmarks typically focus on labeling text with primary emotions like *"happy"* or *"sad"* without requiring the model to understand underlying social context or emotional consequences. As a result, they do not test the model's capacity for deeper emotional comprehension or moral sensitivity.

To address this gap, this thesis utilizes both the GoEmotions [DMAK+20] and the EmoBench benchmark [SLZ+24], the latter being a comprehensive dataset of 400 questions in both English and Chinese, designed specifically to test complex emotional reasoning [SLZ+24]. The dataset GoEmotions is created by Google, and is among the most detailed and frequently used emotion-labeled datasets. It contains more than 58,000 Reddit posts in English tagged with 27 different emotional labels, plus neutral. It represents subtle and mixed emotional expressions, which makes it useful for training and testing emotion recognition models.

Although GoEmotions offers detailed emotional labeling within individual statements, EmoBench takes this further by testing how well models understand emotional context, social behavior patterns, and underlying intentions across multiple scenarios. EmoBench requires models to analyze emotions in complex social settings, evaluate motives, and understand cultural expectations. Earlier evaluations using this benchmark have shown that even the most advanced LLMs regularly fall short when compared to human reasoning. Therefore, EmoBench serves as a useful method to explore the capabilities and weaknesses of DeepSeek R1 in emotional understanding.

DeepSeek-R1 is able to notice simple emotional signals, but it struggles when it comes to understanding complex emotional reasoning. This includes social situations, perspective-taking, and nuanced interpretations. The model is more effective at finding what causes emotions than recognizing the emotions themselves. It also often loses track of the bigger picture throughout changing scenarios. These problems seem to come mainly from its limited experience of the real world, which points to some of the bigger challenges in creating LLMs that truly grasp human emotions.

Thesis Overview: This bachelor thesis begins with an introduction that sets the stage for the research. Chapter 2 provides key definitions relevant to the study, while Chapter 3 reviews existing literature and related work in the field. The methodology and experimental setup are detailed in Chapter 4, followed by the presentation and analysis of results in Chapter 5. Chapter 6 offers an in-depth discussion of the findings, and Chapter 7 summarizes the conclusions and implications of the research. Finally, Chapter 8 addresses the limitations of this work.

This bachelor thesis was completed at the Leiden Institute of Advanced Computer Science (LIACS) under the supervision of Suzan Verberne and Mert Yazan, whose guidance was invaluable throughout the project.

2 Definitions

These definitions are drawn from the EmoBench paper [SLZ+24] and are essential for grasping the concepts and analysis presented in this thesis.

2.1 Complex Emotions

Being able to understand complex emotions is a key part of emotional intelligence. In the EmoBench study [SLZ+24], the researchers of EmoBench focused on three main types of emotional complexity: emotion transitions, emotional mixtures, and unexpected emotional outcomes.

Emotion Transitions People's emotions often change depending on what is happening around them. To test if language models can follow these changes, stories were created by the researchers of EmoBench where a person's emotional state shifts during the situation.

Mixture of Emotions Unlike many past studies that label only one emotion per situation, this benchmark recognizes that people often feel more than one emotion at the same time. These mixed feelings can be similar (like happy and excited) or completely different (like proud and disappointed).

Unexpected Outcome Some emotional reactions do not follow the usual patterns. Therefore, Emobench researchers created scenarios where the emotional response goes against what most people would expect. These are useful for testing whether a language model relies too heavily on typical emotional patterns.

2.2 Personal Beliefs and Experiences

To understand emotions more deeply, it is needed to consider a person's past experiences, cultural background, and values. This section describes how these personal factors can influence emotional reactions.

Cultural Values Different cultures view situations in different ways. To explore this, examples were included where two people have different emotional responses to the same situation based on their cultural norms.

Sentimental Value How much something means to a person can change how they feel when it's lost or damaged. The point is to explore whether the model could recognize when something has emotional or sentimental value.
Example: Losing an old T-shirt you were planning to throw away might not feel like a big deal. But losing a T-shirt that was a gift from a loved one can feel devastating.

Persona A person's personality, fears, or past experiences can shape how they respond emotionally. In this subcategory, it is tested whether the model could adjust its understanding of emotions based on these traits.

2.3 Emotional Cues

Emotional intelligence involves being able to notice and understand emotional signals in both ourselves and other people. While recent studies show that large language models (LLMs) can recognize and respond to direct emotional information, like when emotions are clearly stated, it is still unclear how well they can handle more subtle or hidden emotional clues.

To study this, one of the categories in the EmoBench benchmark focuses on emotional cues that come from how people speak (vocal cues like tone or sighing) and how they act or look (visual cues like facial expressions). These kinds of signals are often not directly described in the text, so understanding them requires deeper emotional reasoning. For example, if someone's face turns red, that could mean they are angry or embarrassed. Or if someone sighs, it might show relief or frustration depending on the context.

2.4 Perspective Taking

Another key part of emotional understanding is the ability to take someone else's point of view. This skill is often called *"perspective-taking"*, and it's also an important part of what researchers call Theory of Mind (ToM), which is the ability to imagine what someone else is thinking or feeling [FF05]. In this benchmark, three common tasks are used from ToM research to test if LLMs can understand emotions by imagining how different people see the same situation. These tasks are: Affective False Belief, Faux Pas, and Strange Story.

Affective False Belief This task is based on the well-known *"Sally-Ann"* test [BCLF85], which is used to see if a person has the skill of understanding what other people might be thinking or feeling based on their situation. LLMs are tested in similar ways by showing them situations where someone does not know the full story. To answer correctly, the model has to reason about what the person *thinks* is true and how that affects their emotional reaction.

Faux Pas A *"faux pas"* is careless comments or actions that accidentally hurt someone or lead to a negative outcome. The faux pas test [BCOS+99], people are given a social scenario and asked to figure out whether someone said or did something socially inappropriate or awkward. In the version of Emobench, researchers focus on how well LLMs can recognize the emotional effects of a faux pas. To do this, the model has to understand not only what was said but also what each person in the situation knows.

Strange Story This test, inspired by [BSS+96], includes stories where things do not follow normal rules. These scenarios are used to see if the model can adjust its emotional reasoning when situations are unusual or unexpected. The idea is to find out if the LLM truly understands the situation or if it's just guessing based on common patterns.

> *Usually, getting an F on a test means failure and disappointment. But in a class where the teacher gives Fs only to the best students, getting an F would make the student feel proud.*

3 Related work

For this research, we draw on three central resources: the DeepSeek Thoughtology study [MPA+25], which offers insights into the architecture and reasoning abilities of the R1 model, and two datasets, GoEmotions [DMAK+20] and EmoBench [SLZ+24]. This section introduces the datasets used to evaluate different aspects of emotional understanding. GoEmotions supports broad emotion classification, while EmoBench challenges the model's ability to reason through complex, socially grounded emotional scenarios.

3.1 DeepSeek Thoughtology Paper

The paper *DeepSeek-R1 Thoughtology: Let's Think About LLM Reasoning* [MPA+25] provides an in-depth examination of the internal reasoning processes of the DeepSeek-R1 model. Given that this model is central to our thesis, it is essential to understand how it operates, particularly how it generates, evaluates, and occasionally overextends its reasoning. One of the most intriguing aspects is its tendency to *"overthink"*, a phenomenon further explored in the Discussion section. Since our research focuses on how large language models (LLMs) perceive and interpret emotions, gaining insight into their reasoning pathways is crucial for evaluating their emotional understanding.

3.1.1 Overthinking and Looping Errors

DeepSeek-R1 exhibits a range of behaviors that can be interpreted as cognitive over-processing, or what might be described as "overthinking." Drawing from prior observations [MPA+25], several recurring patterns were identified that reflect inefficiencies in the model's reasoning process.

1. **Redundant Reasoning:** The model often revisits prior decisions without adding anything meaningful or new.

2. **Excessive Focus on Minor Details:** DeepSeek-R1 focuses on minor details and fails to identify the important information.

3. **Considering Too Many Options:** DeepSeek-R1 engages in overly detailed analysis of too many options.

4. **Lack of Thought Management:** DeepSeek-R1 keeps thinking and explaining even when it is not needed, which makes its answers more complicated and less efficient.

5. **Inconsistent Context Tracking:** DeepSeek-R1 often has trouble following the flow of a situation as it changes over time.

These patterns suggest that while the model is capable of complex reasoning, it does not always apply it effectively. The lack of prioritization and planning within its cognitive sequence often leads to looping behaviors that resemble overthinking in human cognition.

3.2 Emotional Understanding in LLMs

Lately, LLMs have made noticeable progress in cognitive and linguistic capabilities [SLZ+24]. While these systems handle defined tasks effectively, such as summarization and question answering, their capacity for emotional reasoning remains less explored. Emotional intelligence (EI) is more than just basic emotion recognition [BLMVMSCS21]; it involves interpreting, managing, and responding to emotions in a socially and contextually appropriate manner. However, many existing models tend to rely on superficial textual cues rather than engaging in deeper emotional inference [WLY+23]. As LLMs become more advanced, it is becoming more important to have more thoughtful evaluation tools that measure their ability to reason about complex emotional and psychological states. This thesis addresses this need by employing a benchmark specifically developed to assess nuanced emotional understanding and applies it to evaluate the performance of the DeepSeek-R1 model.

3.3 GoEmotions

GoEmotions is one of the most extensive publicly available emotion datasets, comprising over 58,000 English-language Reddit comments. Each comment is annotated with one or more emotion labels drawn from a taxonomy of 27 emotions, along with a neutral category. The labels were assigned through a rigorous manual annotation process, aiming to ensure high data quality and consistency across diverse emotional expressions [DMAK+20].

The GoEmotions dataset serves as a valuable resource for training and benchmarking models on multi-label emotion classification tasks. Its rich diversity of user-generated content makes it particularly effective for assessing a model's ability to identify a wide range of emotional states in everyday language. In previous studies, transfer learning experiments have shown that models trained on GoEmotions generalize reasonably well to other emotion-related tasks and datasets. In this research, GoEmotions is used to examine DeepSeek R1's performance on identifying explicit emotional expressions, providing a baseline before advancing to more complex forms of emotional reasoning.

3.4 EmoBench: A Benchmark for Evaluating Emotional Intelligence in Large Language Models

EmoBench is a recent benchmark designed to evaluate advanced emotional reasoning in LLMs. Unlike basic emotion classification datasets like GoEmotions, EmoBench focuses on deeper emotional understanding and context-based reasoning. It includes 400 carefully crafted multiple-choice questions in English and Chinese, testing models on scenarios involving empathy, regulation, and social dynamics.

This benchmark is particularly valuable for its real-world relevance and multilingual scope. Rather than relying on obvious emotional cues, EmoBench challenges models to interpret subtle implications, making it well-suited for assessing complex emotional intelligence. Its design better reflects how emotions function in everyday life, where meaning often depends on unspoken context and personal perspective.

In this thesis, EmoBench is used to evaluate DeepSeek R1's ability to handle nuanced situations, including those with moral or interpersonal tension. While GPT-4 was used to generate initial prompts, the final dataset was manually refined to increase diversity and emotional depth. This ensures that the benchmark tests reasoning, not just pattern recognition.

The following figures, adapted from the EmoBench paper, show the difference between traditional design and the EmoBench design.

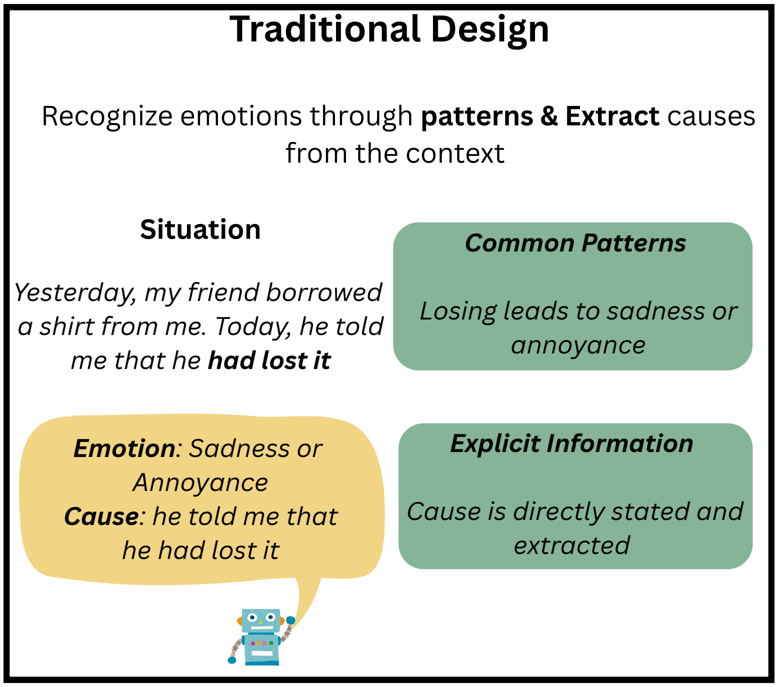

Traditional Design
Recognize emotions through **patterns & Extract** causes from the context
Situation
Yesterday, my friend borrowed a shirt from me. Today, he told me that he ***had lost it***
Common Patterns
Losing leads to sadness or annoyance
***Emotion**: Sadness or Annoyance*
***Cause**: he told me that he had lost it*
Explicit Information
Cause is directly stated and extracted

Figure 1: Traditional Design

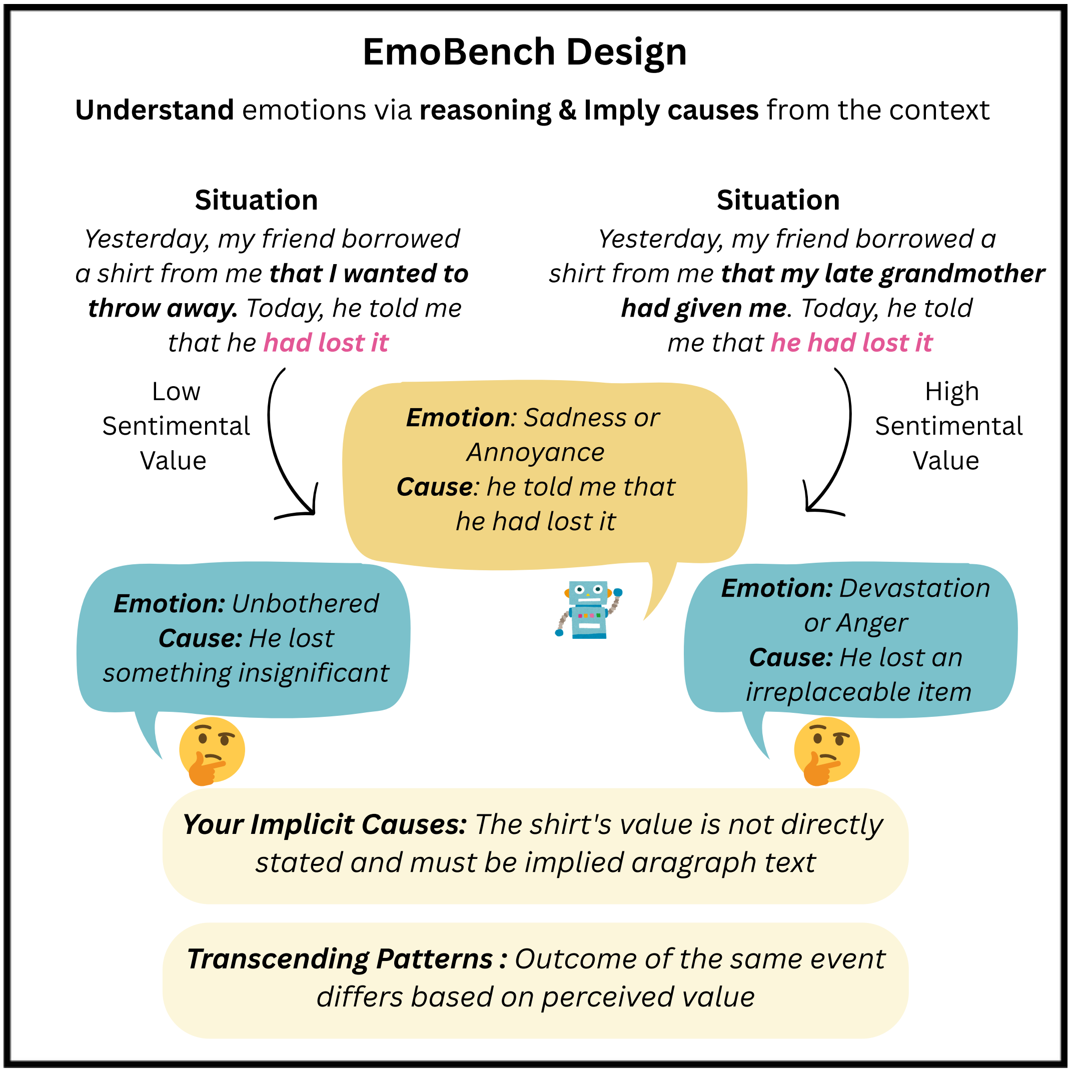

EmoBench Design
Understand emotions via reasoning & Imply causes from the context
Situation
Yesterday, my friend borrowed a shirt from me that I wanted to throw away. Today, he told me that he had lost it
Situation
Yesterday, my friend borrowed a shirt from me that my late grandmother had given me. Today, he told me that he had lost it
Low Sentimental Value
High Sentimental Value
Emotion: Sadness or Annoyance
Cause: he told me that he had lost it
Emotion: Unbothered
Cause: He lost something insignificant
Emotion: Devastation or Anger
Cause: He lost an irreplaceable item
Your Implicit Causes: The shirt's value is not directly stated and must be implied aragraph text
Transcending Patterns : Outcome of the same event differs based on perceived value

Figure 2: EmoBench Design

As illustrated in the figures above, traditional emotion recognition tools often depend on simplistic keyword-emotion pairings. For instance, models may automatically associate the word *lost* with the emotion *sadness*, without accounting for the personal significance of the lost object. EmoBench, however, adds complexity by embedding contextual elements, thus requiring models to evaluate not only the emotional state associated with the word but also the emotional significance of the context. For example, losing a pen might trigger a minimal emotional response, whereas losing a sentimental gift could evoke a much stronger emotional reaction. By incorporating such contextual factors, EmoBench encourages a more nuanced evaluation of emotional intelligence in LLMs.

3.4.1 Benchmark Design

EmoBench evaluates large language models on two key emotional intelligence dimensions: **Emotional Understanding (EU)** and **Emotional Application (EA)**. This thesis emphasizes EU, which tests how well a model can interpret emotionally complex and nuanced situations.

- **Emotional Understanding (EU)**: EU involves reasoning about emotional experiences by considering beliefs, interpersonal dynamics, and subtle cues. EmoBench groups this into four areas: complex emotions, emotional cues, personal beliefs, and perspective-taking (see Section 2). Plutchik's Wheel of Emotions is used to label emotional states and intensities.

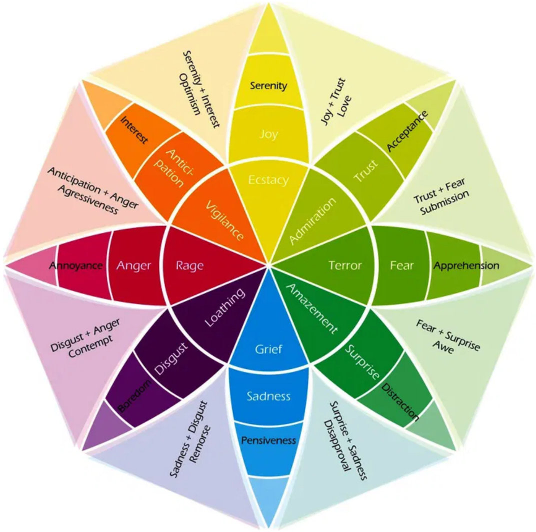


Figure 3: Plutchik's Wheel of Emotions

Researchers created 121 emotionally rich scenarios involving one to three individuals. Each individual's emotional state and cause were annotated, then transformed into multiple-choice questions (MCQs), resulting in 200 total questions. Translations and multi-stage reviews ensured clarity and consistency.

- **Emotional Application (EA)**: EA tests whether a model can apply emotional knowledge to choose contextually appropriate responses in emotionally sensitive situations.

Each EmoBench entry includes the following elements:

- **Category**: The thematic subdomain
- **Scenario**: A bilingual (EN/CH) emotional context
- **Subject**: The central individual in the scenario
- **Emotion**: A list of possible emotional outcomes
- **Label**: The correct emotion based on the scenario
- **Cause**: An explanation for the emotional response

3.4.2 Scenario Design and Answer Annotation

To ensure the model would engage in thoughtful reasoning, each scenario was paired with multiple answer choices, all crafted to appear equally plausible. This design increased the difficulty of the task by requiring nuanced emotional understanding rather than relying on superficial cues. Once the scenarios and answer choices were developed, a second annotator reviewed and translated the material into either English or Chinese, depending on the original language.

Given the inherently subjective nature of emotional interpretation, the annotation process adopted a scoring method based on the framework introduced by MacCann and Roberts [AWH+14]. Each multiple-choice question (MCQ) was evaluated by a total of four annotators, two involved in the original development and two independent reviewers. Annotators were asked to distribute a total score of 1.0 across the available answer options, allowing them to express partial preferences. For instance, if an annotator considered two answers plausible but favored one slightly more, they could assign 0.75 to the preferred option and 0.25 to the other.

To determine the most appropriate response for each scenario, the individual scores from all annotators were averaged. This aggregation method ensured that the final selected answer reflected a balanced human judgment. In order to assess the consistency of the annotations, Fleiss' Kappa was calculated as a measure of inter-annotator agreement [NWCG10]. The resulting value of $\kappa = 0.852$ indicates a high level of agreement among annotators, supporting the reliability of the annotated dataset.

3.4.3 Experiments conducted in the EmoBench paper

Each task is conducted in two settings: a basic setting with zero-shot prompting using only task instructions (Base), and a more advanced setting that encourages step-by-step reasoning before selecting an answer (Chain-of-Thought or CoT). The specific prompts used for both settings are available in the Emobench paper [SLZ+24].
To reduce variance and increase reliability, each model answers every MCQ five times. The researchers determine the final prediction using majority voting across the five outputs. Additionally, since language models can show a bias toward certain answer positions [ZZM+23], we shuffle the answer order in three alternative ways, generating four permutations in total. Each permutation is evaluated separately, and we report the average accuracy across all runs.

3.4.4 Detailed Results and Error Analysis

The evaluation revealed that LLMs encountered considerable difficulty with questions involving *perspective-taking*, particularly those requiring an understanding of emotional states from another individual's point of view. Specific challenges were observed in the following areas:

- Interpretation of personality traits (e.g., distinguishing between shy and assertive behavior)
- Recognition of sentimental or emotional attachment to objects
- Awareness of cultural norms and how behaviors may vary in appropriateness across cultural contexts

Language variation had only a minor impact on model performance. Most models demonstrated slightly better accuracy in English, which may be attributed to an over-representation of English data in their training corpora. However, language-specific models such as Yi outperformed English-centric models in tasks conducted in Chinese, suggesting a training-data advantage for their respective language.

While LLMs do not possess emotional awareness in the human sense, their performance suggests a capacity to simulate empathy and affective understanding through pattern recognition and linguistic inference. This apparent understanding stems not from true emotional experience, but from exposure to extensive text data containing emotional expressions and narratives.

To minimize subjectivity and ensure evaluative consistency, each item in the EmoBench dataset was constructed with a single, clearly correct answer. These answers were selected based on standardized emotion definitions (e.g., sadness, anger, pride) and reviewed by four independent human annotators. The high level of agreement among annotators, reflected by a Fleiss' Kappa score of 0.852, indicates that most questions had a reliably identifiable correct response.

4 Methods

This section outlines the methodology we used to assess the emotional reasoning capabilities of DeepSeek-R1. In this study, we focus on how well DeepSeek can identify emotional states and justify them across varied scenarios. The approach is both quantitative and qualitative, leveraging existing emotion recognition datasets and in-depth error analysis to understand the model's strengths and limitations.

4.1 Research Design

The main goal of this study is to find out why LLMs have trouble understanding real emotions, even though they are good with language. To explore this, we divided the research into several steps:

1. Evaluation of the model using two benchmark datasets: GoEmotions and EmoBench.
2. Analysis of DeepSeek's predictions and justifications, particularly in emotionally complex or context-dependent scenarios.
3. Integration of literature and theoretical frameworks to interpret observed failures and limitations.

4.2 Phase 1: Initial Testing with GoEmotions

To investigate the emotional reasoning capabilities of DeepSeek-R1, two main evaluations were conducted using the GoEmotions dataset and the EmoBench benchmark. These evaluations were designed to assess both the model's emotion recognition abilities and its reasoning about emotional causes within realistic, nuanced scenarios.

The first stage involved the GoEmotions dataset, where the model was assigned to identify the dominant emotion in simple English sentences. This dataset provides a straightforward test of emotional classification and helps establish a baseline understanding of the model's recognition performance.

GoEmotions is a detailed dataset created by Google for fine-grained emotion classification. It contains 58,000 English comments from Reddit, each labeled with one or more of 27 different emotions. For our initial evaluation, we used the training portion of this dataset. Every entry includes a short text and an associated emotion label. This allowed us to test how well DeepSeek-R1 can understand clear emotional signals in fairly simple situations.

We presented each sentence from the GoEmotions training set to DeepSeek-R1 as a prompt, asking it to identify the main emotion expressed. The model's answers were then compared with the original labels to measure its accuracy.

However, the GoEmotions dataset proved to be too simple for our purposes. Many sentences in the dataset were short and lacked the complexity or ambiguity required to assess deeper emotional reasoning. While useful for establishing a baseline, GoEmotions did not provide sufficient challenge for evaluating advanced emotional understanding. Consequently, we chose to use EmoBench, which offers richer, more nuanced scenarios better suited for testing higher-order emotional intelligence.

4.3 Phase 2: Main Evaluation with EmoBench

Because GoEmotions has some limitations, we moved on to using EmoBench, a benchmark created to evaluate language models' emotional understanding, reasoning, and empathy. EmoBench includes various categories that challenge models with more complex emotional situations. It is specifically designed to test how well models handle complicated social and emotional contexts. The benchmark covers subcategories like complex emotions, emotional cues, and perspective-taking. Each task requires not only identifying an emotion but also understanding its cause, which better reflects real human emotional reasoning. We analyzed DeepSeek-R1's answers for both accuracy and the depth of its reasoning.

In both tests, we closely examined how the model behaved, focusing on specific problems such as over-analyzing, misunderstanding context, and struggling with complex emotional signals. These issues were grouped into broader challenges, including reliance on pattern matching, excessive reasoning, and difficulty mimicking human-like emotional experience.

For our study, we focused specifically on the *"Emotional Understanding"* (EU) component of EmoBench. This subset includes multiple-choice questions that require the model to:

1. Identify the final emotional state of a character given a scenario.
2. Justify that emotional state by selecting the appropriate cause or rationale.

Each entry includes a narrative prompt followed by two questions and a set of options. To facilitate batch evaluation, we created a JSON file containing each EU item, the ground truth labels, DeepSeek's predictions, and the model's reasoning.

4.4 Experimental Procedure

Each test item in the EmoBench dataset followed a consistent structure, mirroring the standardized EmoBench format:

- A short narrative scenario describing an event involving one or more characters.
- Question 1: *"What emotion(s) would [subject] ultimately feel in this situation?"*
- Question 2: *"Why would [subject] feel [emotions] in this situation?"*

These questions were posed to DeepSeek-R1 using a Python script that connects to its API interface. The model was given each scenario and set of options and instructed to choose the most appropriate answer for each question. Its responses were logged, including the explanations generated as part of DeepSeek-R1's default chain-of-thought reasoning process.

To keep the evaluation fair and consistent, we used the same prompt format as EmoBench. Each prompt was a multiple-choice question, and the model had to pick one answer. We also used the exact answer choices from EmoBench such that we could directly compare the model's answers to the original labels.

We checked how well the model did by comparing its answers to the correct labels in the EmoBench dataset. Every item in the Emotion Understanding (EU) subset was given as a multiple-choice question to DeepSeek-R1. We calculated accuracy separately for both recognizing emotions and identifying their causes across different categories. This helped us see clearly where the model performed well and where it struggled in understanding emotions and context.

4.5 Qualitative Analysis

Beyond accuracy metrics, we conducted a qualitative error analysis to better understand the reasoning failures behind incorrect model responses. Specifically, we:

- Identified common failure types, such as misattribution of emotion or flawed reasoning.
- We looked at the explanations generated by the model to check if they were consistent and similar to how humans reason.
- Highlighted recurring patterns, including an over-reliance on surface-level linguistic cues or failure to maintain the context in the story.

This analysis allowed us to go beyond surface-level correctness and explore why specific errors occurred. For instance, we noticed that DeepSeek often defaulted to generic emotional responses when faced with ambiguous or nuanced scenarios.
A common pattern across models was that they often analyzed the answer choices themselves rather than deeply engaging with the situation. This strategy helped eliminate obviously incorrect options, but it frequently missed the deeper emotional understanding that humans use when making decisions.

5 Results

This section presents both quantitative outcomes and qualitative insights gained from evaluating DeepSeek-R1.

5.1 EmoBench Evaluation: Quantitative Results

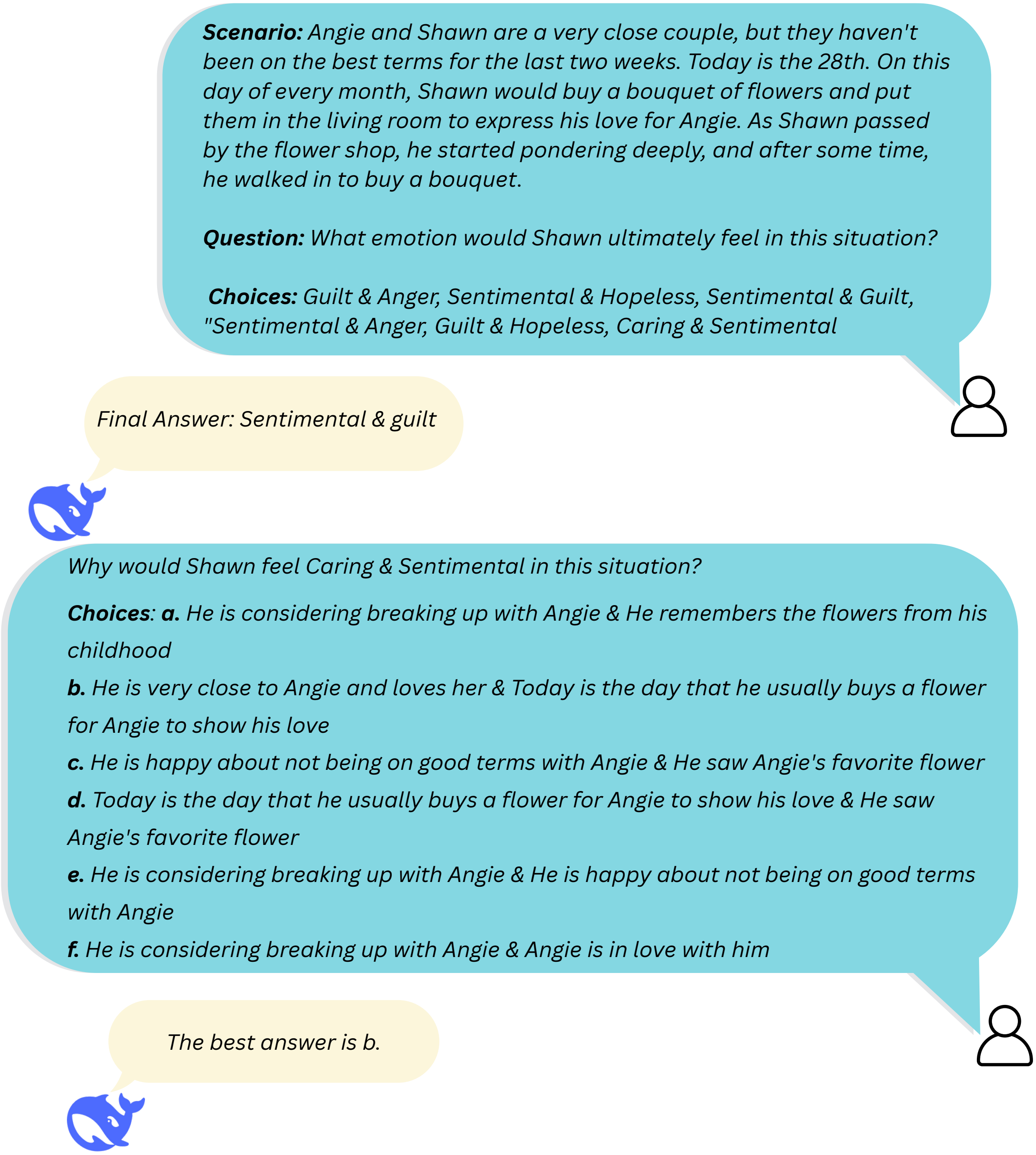
Table 1: Emotional Understanding (EU) Accuracy (%) on EmoBench: DeepSeek-R1 vs. GPT-4

Category	**Task**	**DeepSeek-R1**	**GPT-4 (CoT)**
Complex Emotions	Emotion	75.5	68.9 (EN)
Complex Emotions	Cause	73.5	—
Personal Beliefs	Emotion	55.4	53.1 (EN)
Personal Beliefs	Cause	82.1	—
Emotional Cues	Emotion	64.3	71.4 (EN)
Emotional Cues	Cause	89.3	—
Perspective Taking	Emotion	43.3	49.3 (EN)
Perspective Taking	Cause	88.1	—
Overall	Emotion	57.5	58.3 (EN)
Overall	Cause	83.0	—
EU Total	Both	70.3	—

In this table, it is evident that DeepSeek performs substantially better at identifying emotional causes than emotional states. While it achieves an overall accuracy of 83.0% on cause identification, its performance drops to 57.5% when asked to determine the correct emotional outcome. For GPT-4, it scores an accuracy of 58.3% for determining the correct emotional outcome.

Perspective Taking emerged as the most challenging subcategory. With only 43.3% accuracy in identifying emotions, the model struggled to infer how characters might feel based on another person's viewpoint. Conversely, it performed well in identifying the causes in these same scenarios, with an accuracy of 88.1%. This disparity suggests a disconnect between DeepSeek's ability to understand factual context and its ability to simulate emotional inference.

5.2 Example Case Analysis



Scenario: *Angie and Shawn are a very close couple, but they haven't been on the best terms for the last two weeks. Today is the 28th. On this day of every month, Shawn would buy a bouquet of flowers and put them in the living room to express his love for Angie. As Shawn passed by the flower shop, he started pondering deeply, and after some time, he walked in to buy a bouquet.*

Question: *What emotion would Shawn ultimately feel in this situation?*

Choices: *Guilt & Anger, Sentimental & Hopeless, Sentimental & Guilt, "Sentimental & Anger, Guilt & Hopeless, Caring & Sentimental*

Final Answer: Sentimental & guilt

Why would Shawn feel Caring & Sentimental in this situation?

Choices**:* ***a. *He is considering breaking up with Angie & He remembers the flowers from his childhood*

b. *He is very close to Angie and loves her & Today is the day that he usually buys a flower for Angie to show his love*

c. *He is happy about not being on good terms with Angie & He saw Angie's favorite flower*

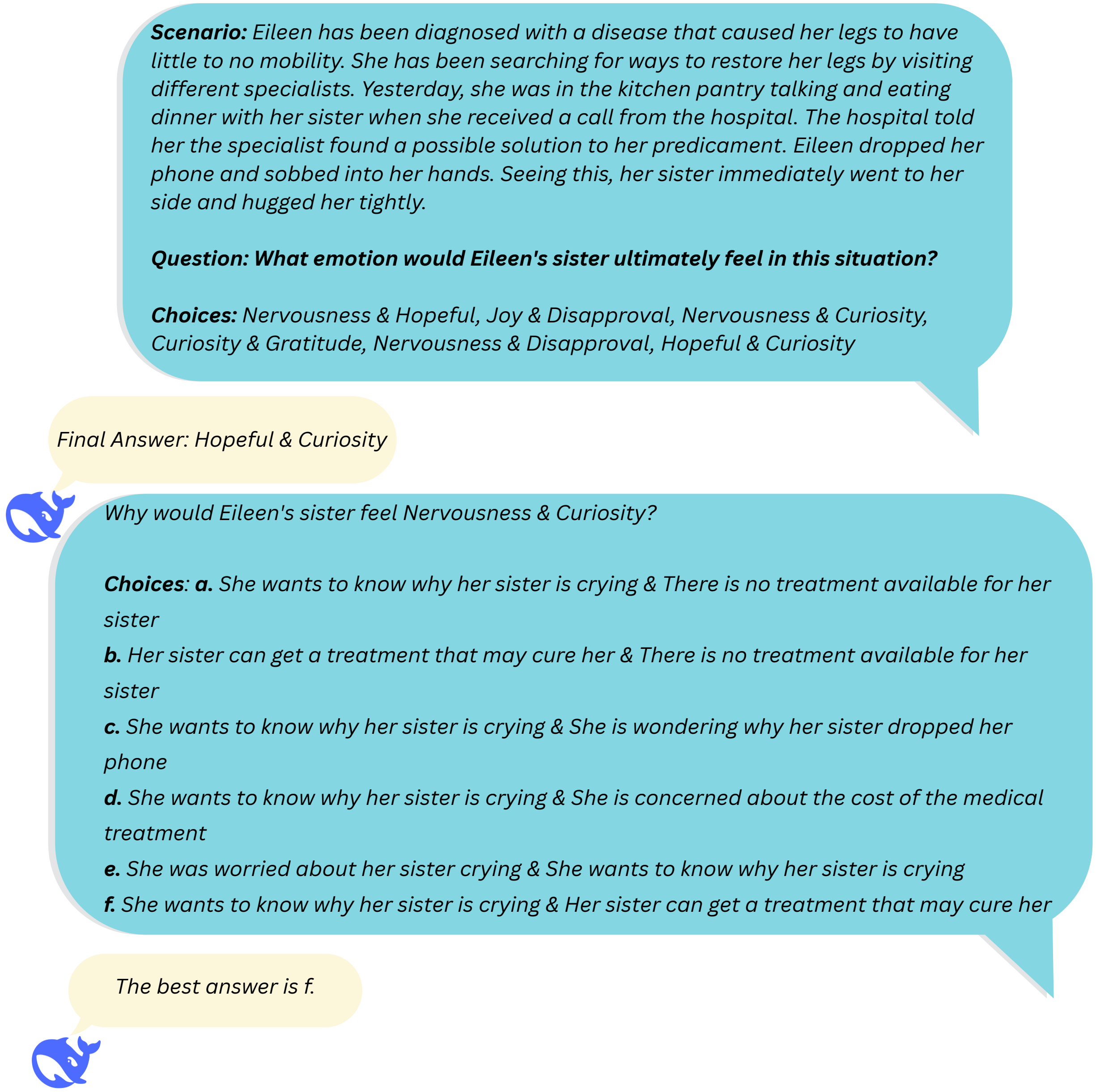
d. *Today is the day that he usually buys a flower for Angie to show his love & He saw Angie's favorite flower*

e. *He is considering breaking up with Angie & He is happy about not being on good terms with Angie*

f. *He is considering breaking up with Angie & Angie is in love with him*

The best answer is b.

In this instance, DeepSeek was unable to correctly identify the emotion (Caring & Sentimental) but accurately determined the cause. This result highlights several potential reasons for DeepSeek's failure, which we will analyze based on previous research. These include its heavy dependence on patterns rather than real-world experience, its tendency to overthink, difficulty with realistic human reasoning, and challenges in maintaining context over time. In this study, we will primarily focus on examining the model's difficulties with realistic human reasoning. These limitations will be examined in greater detail in the subsequent discussion section. Other contributing factors will be briefly acknowledged; however, an in-depth discussion is beyond the scope of this study.



Scenario: *Eileen has been diagnosed with a disease that caused her legs to have little to no mobility. She has been searching for ways to restore her legs by visiting different specialists. Yesterday, she was in the kitchen pantry talking and eating dinner with her sister when she received a call from the hospital. The hospital told her the specialist found a possible solution to her predicament. Eileen dropped her phone and sobbed into her hands. Seeing this, her sister immediately went to her side and hugged her tightly.*

Question: What emotion would Eileen's sister ultimately feel in this situation?

Choices: *Nervousness & Hopeful, Joy & Disapproval, Nervousness & Curiosity, Curiosity & Gratitude, Nervousness & Disapproval, Hopeful & Curiosity*

Final Answer: Hopeful & Curiosity

Why would Eileen's sister feel Nervousness & Curiosity?

***Choices**: **a.** She wants to know why her sister is crying & There is no treatment available for her sister*
b. *Her sister can get a treatment that may cure her & There is no treatment available for her sister*
c. *She wants to know why her sister is crying & She is wondering why her sister dropped her phone*
d. *She wants to know why her sister is crying & She is concerned about the cost of the medical treatment*
e. *She was worried about her sister crying & She wants to know why her sister is crying*
f. *She wants to know why her sister is crying & Her sister can get a treatment that may cure her*

The best answer is f.

In this case, DeepSeek failed to identify the correct emotion (Nervousness & Curiosity) and its cause. The correct cause is e. DeepSeek's answer assumes the sister already knows the content of the call, which is not the case.

6 Discussion

6.1 Heavy Dependence on Patterns Rather Than Experience

Large Language Models, such as DeepSeek, base their outputs on patterns and regularities extracted from their training data. Rather than engaging with the specific context of a scenario, these models often default to emotionally or causally plausible responses that align with common patterns found in language use. This pattern limits the model's ability to recognize subtle emotions or respond appropriately to emotions that change depending on the situation [LBZ+24].

While LLMs are exposed to vast textual data containing a wide range of emotional expressions and contextual indicators, their understanding remains fundamentally probabilistic. They form associations between particular linguistic structures and emotional states; for instance, linking phrases like *"I'm thrilled about the results"* with positive affect, or *"This is so frustrating"* with negative sentiment. However, such associations do not reflect genuine comprehension. Instead, they represent learned correlations, not grounded in subjective awareness or experiential understanding of emotion. As a result, the model's predictions may appear contextually relevant but are ultimately based on surface-level linguistic patterns rather than deeper, experiential insight.

The example in Section 5.2, involving Angie and Shawn, DeepSeek seems to interpret Shawn's *"pondering deeply"* as a typical pattern of emotional conflict, which it associates with guilt. However, this misses the broader context of the scenario, which involves a tradition of affection, namely buying flowers, which is not connected to guilt at all. Instead, it indicates a form of care and sentimentality.

6.2 Examples of observed issues:

- Models occasionally went off-topic, shifting into broad reflections about empathy or morality instead of focusing on the specifics of the scenario.
- Step-by-step reasoning sometimes caused the models to drift away from the question entirely.

Several recurring error types were observed, particularly in Emotion Understanding (EU) tasks. These included:

- **Faulty assumptions:** For example, assuming that someone walking into a room would immediately grasp the emotional atmosphere.

  Scenario: After countless rejections, Susan finally received a job offer and began to cry. At that moment, her mother walked into the room and quickly embraced her upon seeing her in tears.

  Question: How does Susan's mother feel?

  Options: (a) Relief, (b) Nervousness, (c) Anger, (d) Delight
  LLM Prediction: (d) Delight → "Her daughter was finally able to get a job offer."
  Correct Answer: (b) Nervousness

The model incorrectly assumes that Susan's mom instantly understands the situation upon entering the room. In reality, she sees her daughter crying without any context, which likely triggers concern, not immediate joy, leading to a misinterpretation of her emotional response. Also, the example in Section 5.2involving Eileen's sister illustrates how DeepSeek assumes the sister already knows the content of the call, which is not the case.

- **Incorrect reasoning:** Like believing that a phobia always leads to fear, or that receiving an "F" is a failure even when it might be the highest possible grade in that situation.

These errors appear to stem from a lack of true emotional understanding. LLMs often failed at perspective-taking and relied heavily on common textual patterns rather than context-specific reasoning. They also tended to overlook unspoken social rules that are intuitive to humans.

Why does this happen? Based on our analysis, we identified several underlying causes:

1. LLMs rely heavily on statistical patterns from their training data, rather than grounded human experience.

2. They struggle with realistic reasoning, the kind that involves tracking a person's emotional state over time, understanding implicit social rules, or adapting to subtle shifts in context.

3. Rather than reasoning from *"what the person in the story knows"*, they often reason based on a generalized view of the situation.

4. DeepSeek, in particular, had difficulty maintaining narrative continuity. It tended to focus on isolated emotional components (e.g., analyzing individual words or phrases) instead of understanding how emotions develop across a full situation.

These findings led us to a deeper investigation into why LLMs struggle with realistic reasoning.

6.3 Difficulty with Realistic Human Reasoning

Human reasoning about emotions and social situations is fundamentally grounded in embodied experience. Individuals interpret emotional cues not merely through language, but by integrating sensory and physical interactions acquired over time. This embodied knowledge allows for a more intuitive and context-sensitive understanding of both literal and figurative language. For example, encountering the phrase *"climbing a mountain"* may trigger a mental simulation of exertion, environmental difficulty, and the emotional significance of persistence or accomplishment. Such simulations are informed by lived, bodily experiences that reinforce abstract concepts and emotional meanings.

In contrast, large language models (LLMs) such as DeepSeek-R1 are trained exclusively on text and lack access to sensory or motor experiences. Their understanding is derived from statistical co-occurrences within textual data, rather than from interaction with the physical world. As a result, they often struggle to interpret metaphors, idiomatic expressions, and socially nuanced language, where meaning typically depends on embodied intuition rather than surface-level text patterns [TSH+20].

This gap between human and machine cognition is further evident in tasks involving social reasoning. A key challenge for LLMs is their lack of Theory of Mind, which is the cognitive ability to attribute mental states such as beliefs, intentions, and emotions to others [Kos24]. This capacity is central to human social cognition and enables individuals to make accurate inferences in emotionally complex or context-dependent scenarios. LLMs, however, do not maintain internal models of other agents and instead rely on textual patterns to approximate reasoning. Consequently, they often misinterpret subtle emotional or social cues that a human would intuitively grasp.

Additionally, the absence of real-world interaction limits the structural coherence of the conceptual frameworks developed by LLMs. While humans acquire language through continuous interaction with their environment, blending perception, action, and reflection, LLMs operate by detecting probabilistic patterns across large-scale corpora. For example, abstract concepts like *"balance"* are experienced physically and repeatedly by humans, leading to robust mental representations. LLMs, lacking this physical referent, form less grounded and more fragmented interpretations. As a result, their output may be linguistically fluent yet deficient in contextual depth and flexibility [Dov24].

6.3.1 Humans and Emotional Interpretation

Humans have an innate ability to perceive and interpret emotions based on lived experiences, which enables them to respond with nuance and empathy in social interactions. This emotional comprehension is deeply grounded in personal experience and sensory input, allowing individuals to understand and react appropriately to complex emotional cues in varying contexts.

In contrast, LLMs lack true emotional awareness. Instead, they analyze textual patterns derived from vast amounts of data, but without the depth of understanding that human emotional experience provides. This results in a superficial processing of emotional content, where LLMs may generate responses that seem contextually relevant on the surface but fail to capture the emotional subtleties of a given situation. For example, even advanced models like GPT-4 have shown difficulty in tasks requiring emotional intelligence, as shown in the table in the Results Section 5 [SLFC23].

The example in Section 5.2, involving Angie and Shawn, illustrates difficulty in emotional interpretation by DeepSeek-R1. DeepSeek focuses too much on Shawn feeling guilty and doesn't pay enough attention to how caring he is by actually buying the flowers. This means the model has trouble connecting what Shawn does with how he feels inside.
Also, people understand that emotions change over time; from feeling unsure or upset to feeling caring and hopeful. But DeepSeek sees these emotions as separate and doesn't follow how they develop, which shows it struggles to understand how emotions progress.

6.3.2 Theory of Mind and Social Reasoning:

Humans possess a well-developed Theory of Mind (ToM), which allows them to infer and attribute mental states, such as beliefs, desires, and intentions, to others. This ability is crucial for predicting behavior and engaging in meaningful social interactions. The ToM helps humans navigate complex social dynamics, from interpreting non-verbal cues to understanding indirect communication.

LLMs, by comparison, exhibit very limited ToM capabilities. Although models like GPT-4 have made substantial advancements in language processing, they perform well below human levels when tasked with understanding and predicting the mental states of others. For instance, research indicates that GPT-4 achieved only 60% accuracy in specific ToM tasks, illustrating the significant gap between human social reasoning and current machine capabilities [VMF+24]. This limitation underscores the challenges faced by LLMs in simulating the depth of human social cognition

The example in Section 5.2, involving Angie and Shawn, illustrates how DeepSeek lacks Theory of Mind. Humans use Theory of Mind to infer what someone like Shawn likely feels. A human might reason: *"He's doing something kind. He must still care, even if things have been hard."* DeepSeek lacks that layered reasoning. It does not simulate Shawn's motivations or emotional goals, instead relying on surface textual cues (e.g., *"pondering deeply"*) without grounding them in intentionality or social norms.

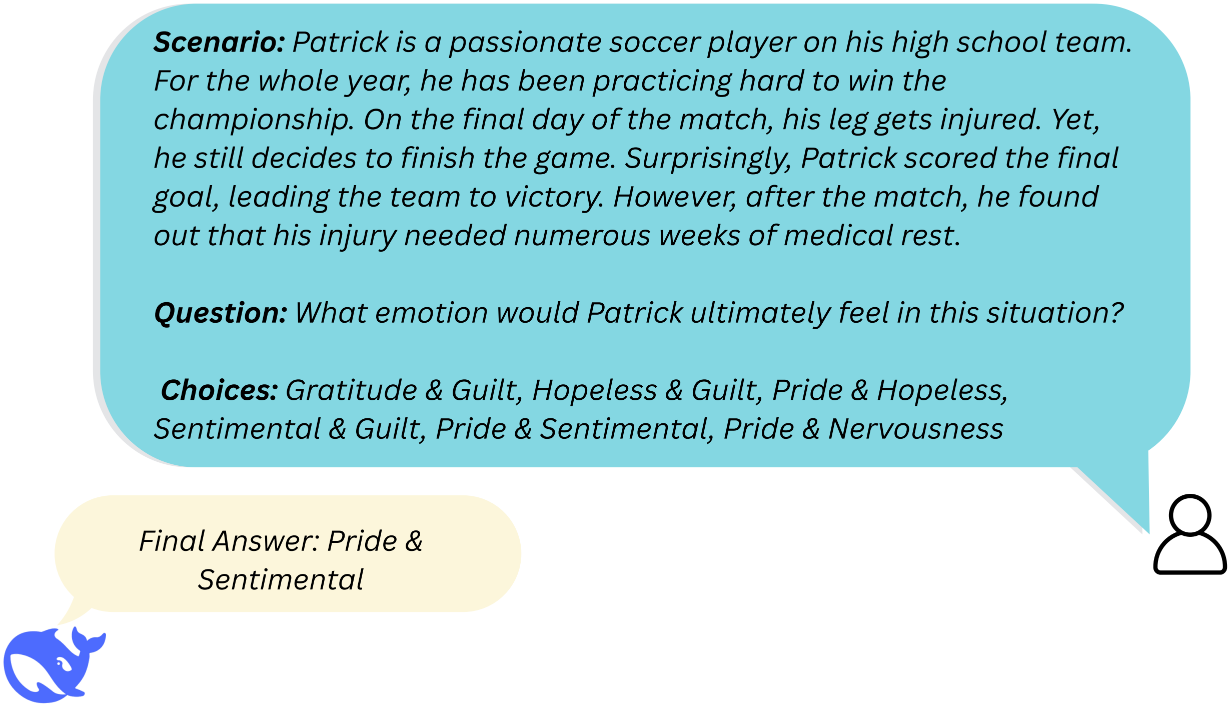
6.3.3 Contextual Understanding:

Humans excel at interpreting nuanced contexts, including sarcasm, idiomatic expressions, and cultural references. This ability to navigate subtle linguistic cues and adapt responses accordingly is a key component of effective communication. Humans can infer meaning from the broader context of a conversation, making it possible to understand jokes, irony, and implied meanings without explicit clarification.

LLMs, on the other hand, often struggle with deeper contextual understanding. While these models can maintain coherence within a conversation and generate text that appears relevant, they are prone to misinterpreting or overlooking subtleties in complex situations. As a result, LLMs may produce responses that are contextually inappropriate or fail to grasp the underlying significance of a message, particularly when it involves indirect communication or culturally specific references.

In summary, the comparison between humans and LLMs highlights critical differences in emotional reasoning, social cognition, ethical understanding, and contextual interpretation. While LLMs represent a significant advancement in natural language processing, their ability to replicate human-like understanding remains limited, particularly in domains requiring emotional depth, social intelligence, and ethical awareness. These gaps illustrate the fundamental challenge of developing machines that can truly mimic the complex, context-sensitive reasoning exhibited by humans.

To illustrate DeepSeek-R1's ability to perform emotional reasoning, an example is presented in the figure below.



***Scenario:** Patrick is a passionate soccer player on his high school team. For the whole year, he has been practicing hard to win the championship. On the final day of the match, his leg gets injured. Yet, he still decides to finish the game. Surprisingly, Patrick scored the final goal, leading the team to victory. However, after the match, he found out that his injury needed numerous weeks of medical rest.*

***Question:** What emotion would Patrick ultimately feel in this situation?*

***Choices:** Gratitude & Guilt, Hopeless & Guilt, Pride & Hopeless, Sentimental & Guilt, Pride & Sentimental, Pride & Nervousness*

Final Answer: Pride & Sentimental

According to EmoBench researchers, the most plausible emotional combination is *Pride & Hopeless*, reflecting the bittersweet nature of the victory, Patrick experiences pride in his achievement, but also a sense of despair over the consequences of his injury. However, DeepSeek-R1 selected *Pride & Sentimental* after extensive internal reasoning, effectively rejecting the more psychologically realistic answer.

This output highlights a notable limitation in the model's reasoning. While it succeeds in recognizing positive emotions like pride and the emotional tone of the event, it underestimates or misclassifies more complex and potentially contradictory feelings such as hopelessness.

> "Hopeless implies no hope, which might not be accurate since it's temporary." DeepSeek-R1

This quote illustrates a key issue: the model equates *hopelessness* with permanent despair and fails to account for situational hopelessness, a temporary but realistic emotional response to a long-term injury following a peak moment. This misjudgment is a form of **difficulty with realistic emotional reasoning**, especially in edge cases involving emotionally mixed outcomes.

Thus, this case study underscores the need for improved affective modeling in LLMs, particularly in parsing emotions that are dynamic, conflicting, or dependent on deeper psychological context.

7 Conclusion

In this study, we set out to explore the research question: **How effectively does the DeepSeek-R1 model understand emotions?** To answer this, we looked at how the model handles both recognizing emotional content and reasoning about emotions in various contexts. Our findings provide insights into DeepSeek-R1's strengths and weaknesses, which show its ability to identify emotional cues as well as its challenges in deeper emotional reasoning.

Evaluations using the GoEmotions dataset and the EmoBench benchmark highlighted a significant limitation: DeepSeek-R1 often fails to fully match human emotional understanding, especially in situations involving subtle or complex emotions. While the model performs well on basic tasks such as identifying obvious emotional expressions, it struggles with complex tasks. These include tasks such as tracking emotions that evolve over time or perspective-taking. This reflects the current limitations of LLMs in matching the emotional depth humans have.

Our results further expose an imbalance in DeepSeek-R1's emotional processing: the model performs significantly better at identifying the causes of emotions than at accurately recognizing the emotions themselves. This suggests that the model lacks deeper contextual and empathetic insight necessary to fully interpret emotional experiences in human terms.

Our results show that the DeepSeek-R1 model struggles with emotional reasoning. It often relies on patterns rather than experience, focuses too much on minor details, struggles to keep track of the overall story, which leads to faulty assumptions and incorrect reasoning.

These limitations mainly come from the fact that the model has no real-life experience or connection to the physical world. Humans learn emotional understanding through years of sensory and social experiences, while DeepSeek-R1 learns only from textual data. Because of this, it struggles to understand abstract ideas, metaphors, or culturally specific ways of expressing emotions. This points to an important area for future improvement in emotional AI.

8 Limitations

While our evaluation of DeepSeek-R1 provided valuable and meaningful insights into how the model understands and reasons about emotions, there are some limitations to keep in mind. These factors might help further studies.

8.1 Cultural and Linguistic Constraints

Human emotions are strongly influenced by culture, and people often express emotions in ways that reflect their social and cultural background. However, LLMs are mostly trained on text from certain languages and cultures. As a result, it may show bias. Thus, the model's answers might not work well for people from different backgrounds, because emotional responses vary by situation and culture.

8.2 Limited Multimodal Input

Another limitation of the current study is its exclusive focus on textual input. In real life, understanding emotions relies on non-verbal cues such as facial expressions, tone of voice, and body language. Since these non-verbal cues were not part of our thesis, the model's emotional understanding ability is limited compared to individuals.

8.3 Subjectivity and Benchmark Design

Understanding emotions often involves personal interpretation. While we tried to reduce bias in the EmoBench benchmark, the task still has some subjectivity. Each question in the Emotional Understanding (EU) section included one answer that was considered correct, based on a set emotion framework. Four different reviewers checked the questions, and their strong agreement shows good consistency. Still, some emotional judgments are hard to standardize fully, even with careful design.

8.4 Individual Differences Not Addressed

In this study, we did not examine how personal factors, such as personality, life experience, or communication style, might affect emotional perception and expression. Emotional reactions can vary significantly between individuals; for example, one person might find a comment funny, while another may find it offensive. While EmoBench scenarios were designed to be broadly interpretable, they still do not capture the full spectrum of human emotional diversity. Future research could explore how such differences affect emotional interpretation.

References

[AWH$^+$14] Veleka D. Allen, Alexander Weissman, Susan Hellwig, Carolyn MacCann, and Richard D. Roberts. Development of the situational test of emotional understanding – brief (steu-b) using item response theory. *Personality and Individual Differences*, 65:3–7, 2014. Emotional intelligence: Research and Applications.

[BCLF85] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46, 1985.

[BCOS$^+$99] Simon Baron-Cohen, Michelle O'riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. 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, 1999.

[BHV$^+$23] Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, 2023.

[BLMVMSCS21] L. M. Bru-Luna, M. Martí-Vilar, C. Merino-Soto, and J. L. Cervera-Santiago. Emotional intelligence measures: A systematic review. *Healthcare*, 9(12):1696, 2021.

[BLS$^+$25] Antonin Brun, Ruying Liu, Aryan Shukla, Frances Watson, and Jonathan Gratch. Exploring emotion-sensitive llm-based conversational ai, 2025.

[BSS$^+$96] James R. Blair, Carol Sellars, Ian Strickland, Fiona Clark, Akintude Williams, Margaret Smith, and Lawrence Jones. Theory of mind in the psychopath. *Journal of Forensic Psychiatry*, 7:15–25, 1996.

[DMAK$^+$20] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions, 2020.

[Dov24] Guy Dove. Symbol ungrounding: what the successes (and failures) of large language models reveal about human cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1911):20230149, 2024.

[FF05] Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.

[Kos24] Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), October 2024.

[LBZ$^+$24] Yujie Li, Erwan Benchetrit, Lingkai Zheng, Chunting Zhang, Xiang Zhang, Chris Kedzie, Debanjan Ghosh, Aaron Steven White, Noah A. Smith, and Varun Gangal. Emobench: Benchmarking emotional intelligence in large language models. *arXiv preprint arXiv:2504.07128*, 2024.

[MPA$^+$25] Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, and Siva Reddy. Deepseek-r1 thoughtology: Let's ¡think¿ about llm reasoning, 2025.

[NWCG10] Thomas R Nichols, Paola M Wisner, Gary Cripe, and Lakshmi Gulabchand. Putting the kappa statistic to use. *The Quality Assurance Journal*, 13(3-4):57–61, 2010.

[SLFC23] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms, 2023.

[SLZ$^+$24] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*, 2024.

[TSH$^+$20] Ronen Tamari, Chen Shani, Tom Hope, Miriam R. L. Petruck, Omri Abend, and Dafna Shahaf. Language (re)modelling: Towards embodied language understanding, 2020.

[VMF$^+$24] Anvesh Rao Vijjini, Rakesh R. Menon, Jiayi Fu, Shashank Srivastava, and Snigdha Chaturvedi. Socialgaze: Improving the integration of human social norms in large language models, 2024.

[WLY+23] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958, 2023.

[XPN+23] Qihui Xu, Yingying Peng, Samuel A. Nastase, Martin Chodorow, Minghua Wu, and Ping Li. Does conceptual representation require embodiment? insights from large language models, 2023.

[ZZM+23] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*, 2023.