Universiteit Leiden
The Netherlands

Data Science and AI

The Effect of Emotional Explicitness on Cognitive and Affective Empathy in Large Language Model Responses

Emine Melisa Uzun

Supervisors:
Max van Duijn & Dorota Moravcikova

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/07/2026

Abstract

This thesis examines whether emotional explicitness in user prompts influences perceived cognitive and affective empathy in large language model responses. A 2×2 within-scenario design was used to compare OLMo-2 7B Instruct and OLMo-3 7B Instruct across implicit and explicit versions of the same emotional scenarios. Human evaluators rated the generated responses on separate cognitive and affective empathy dimensions using an EPITOME-inspired rating procedure. The results showed that explicit emotion labels led to higher perceived empathy ratings for both cognitive and affective empathy. OLMo-3 also received higher empathy ratings than OLMo-2 across both dimensions. In addition, OLMo-2 showed a significant gap between cognitive and affective empathy, while OLMo-3 showed a more balanced pattern. These findings suggest that both prompt wording and model version can influence how empathic LLM-generated responses are perceived to be.

Contents

1 Introduction

Large language models are increasingly used in situations where users describe personal or emotionally sensitive experiences. In these situations, a response may need to recognize the user's emotional state and respond in a way that appears appropriate and supportive. This makes perceived empathy an important part of evaluating LLM-generated responses.

The importance of empathy can be understood through its connection to emotional intelligence. Salovey and Mayer describe emotional intelligence as the ability to understand emotional information in oneself and others and to use this information to guide thought and behaviour. They also present empathy as a central feature of emotionally intelligent behaviour [SM90]. This connection makes empathy relevant for studying whether LLM responses appear emotionally understanding and supportive.

Empathy should not be treated as a single ability. Empathy research commonly distinguishes between cognitive empathy and affective empathy. Cognitive empathy refers to understanding another person's feelings or emotional situation, whereas affective empathy refers to emotionally responding to another person's experience [CBTH16, DST14]. This distinction is important for LLM evaluation, because a response may identify what a user is feeling without expressing much emotional warmth. Conversely, a response may sound warm and supportive without clearly understanding the user's emotional state.

Recent LLM empathy studies have also used this distinction. Yu et al. evaluated cognitive and affective empathy separately when comparing LLMs with human participants [YPF+25]. Liu et al. used separate dimensions in a scenario-based evaluation of LLM responses, linking emotional reactions to emotional empathy and interpretation to cognitive empathy [LCHW25]. These studies show why empathy in LLM responses should be evaluated in more than one dimension.

Other work has shown that LLM-generated responses can be perceived as empathic by human evaluators [LSZ+24]. However, empathy in LLMs remains difficult to evaluate. In their systematic review, Sorin et al. show that existing studies examine LLM empathy across different settings, with a strong focus on healthcare-related contexts. They also indicate that the methods used to evaluate empathy in LLMs remain inconsistent and require further development [SBB+24]. Therefore, this thesis focuses on perceived empathy in generated responses, rather than on whether the model has genuine emotional experience.

The present study focuses on emotional explicitness in user prompts. In real interactions, people do not always state their emotions directly. A user may describe a situation that implies guilt, sadness, happiness, or anxiety without naming the emotion. In other cases, the user may explicitly label the emotion, for example by saying "I felt really content" or "I feel really anxious." To test whether this difference matters, this study compares paired implicit and explicit versions of the same emotional scenarios in responses generated by OLMo-2 7B Instruct and OLMo-3 7B Instruct. This design keeps the scenario content constant while varying whether the emotion is directly named. The generated responses are evaluated separately on cognitive and affective empathy, allowing the study to examine both the effect of emotional explicitness and possible differences between the two model versions.

1.1 Thesis overview

The remaining sections of this thesis are organized as follows. Section 2 introduces the theoretical background and related work, including perceived empathy, Theory of Mind, cognitive and affective empathy, the EPITOME framework, empirical work on LLM empathy, the research gap, and the relevance of analyzing emotional explicitness in LLM-generated responses. Section 3 presents the research questions addressed in this study. Section 4 describes the experimental design, scenario materials, response generation procedure, survey design, data cleaning, and statistical analysis plan. Section 5 reports the results, including inter-rater reliability, descriptive statistics, repeated-measures ANOVAs, and paired-samples t-tests. Section 6 discusses the interpretation of the findings and selected response examples. Section 7 presents the limitations and directions for future research. Finally, Section 8 concludes the thesis.

This bachelor thesis was conducted at the Leiden Institute of Advanced Computer Science (LIACS) under the supervision of Max van Duijn and Dorota Moravcikova. Its main contribution is a controlled evaluation of how implicit and explicit emotional context in user prompts affects perceived empathy in open-weight LLM responses. By comparing OLMo-2 7B Instruct and OLMo-3 7B Instruct, the thesis also examines whether perceived cognitive and affective empathy differs between the two model versions.

2 Background and Related Work

2.1 Empathy and Theory of Mind

Empathy is a central concept in social and psychological research, but it has been defined in different ways. Cuff et al. show that empathy is often difficult to separate from related concepts, such as sympathy, compassion, perspective taking, and emotional contagion [CBTH16]. This is important because empathy is not a single-dimensional ability. It involves several processes, including understanding another person's emotional state, responding to that state, and keeping a distinction between one's own emotions and the emotions of the other person.

This self-other distinction is an important part of empathy. Without this distinction, an emotional response may be closer to emotional contagion, where one person's emotion is simply mirrored by another person. By contrast, empathy requires the observer to understand that the emotion belongs to someone else. This is why empathy is often discussed as both an emotional and a cognitive process. A person may resonate with another person's feelings, but also needs to understand the meaning of those feelings [CBTH16].

A related concept is Theory of Mind. Dvash and Shamay-Tsoory describe Theory of Mind, also called mentalizing, as the attribution of mental states to other people, which allows a person to understand or predict another person's behaviour [DST14]. They also explain that this ability is important for social interaction, but that inferring another person's thoughts and feelings is not the same as empathy. Empathy also involves experiencing and responding to another person's emotional state in a caring or supportive way [DST14]. This makes Theory of Mind especially relevant to the cognitive side of empathy. In the model described by Dvash and Shamay-Tsoory, cognitive empathy is connected to Theory of Mind processes, including cognitive ToM and affective ToM, whereas emotional empathy is more closely linked to simulation and emotional experience [DST14]. This distinction helps explain why empathy can involve both understanding another person's mental or

emotional state and emotionally responding to that state.
This relationship is also reflected in the work of Preckel, Kanske, and Singer. They distinguish between socio-cognitive and socio-affective routes to understanding others [PKS18]. Theory of Mind is mainly related to the socio-cognitive route, because it concerns understanding another person's thoughts and intentions. Empathy is more closely related to the socio-affective route, because it concerns resonance with another person's feelings. These processes are related, but they can also differ. For example, someone may understand what another person is thinking without strongly sharing their feelings. Conversely, someone may be emotionally affected by another person's distress without accurately interpreting their mental state.
Although empathy and Theory of Mind are distinct, they often work together in social situations. Understanding another person's experience may require perspective taking, reasoning about their situation, and responding to their emotional state. This is especially relevant when emotions are not stated explicitly. In such cases, the observer has to infer the person's emotional state from contextual cues.
Overall, empathy and Theory of Mind should be treated as connected but separate concepts. Theory of Mind is especially relevant to the cognitive dimension of empathy, because it concerns the interpretation of another person's mental or emotional state. However, empathy also includes an affective dimension that goes beyond mental-state reasoning. This distinction provides the theoretical basis for separating cognitive and affective empathy in the next subsection.

2.2 Cognitive Empathy and Affective Empathy

Empathy research commonly separates cognitive empathy from affective empathy. Cognitive empathy refers to understanding another person's feelings, perspective, or emotional situation. It concerns the interpretative side of empathy: recognizing what another person may be experiencing and making sense of that experience from their point of view. Cuff et al. describe cognitive empathy as understanding another person's feelings and note that empathy is often discussed as involving both cognitive and affective components [CBTH16]. In the rating procedure used in this study, this dimension captures whether a response appears to understand the user's emotional situation.
Affective empathy refers to the emotional response to another person's experience. It concerns emotional resonance, sharing, or concern in response to another person's emotional state. Cuff et al. describe affective empathy as the experience of emotion in response to an emotional stimulus [CBTH16]. Dvash and Shamay-Tsoory similarly distinguish between a cognitive empathic system and an emotional empathic system, where the emotional system is more closely related to simulation and emotional experience [DST14]. For the present study, this dimension is reflected in the warmth, care, compassion, and emotional support expressed in a response.
Although cognitive and affective empathy are related, they should not be treated as identical. Dvash and Shamay-Tsoory argue that cognitive and emotional empathy are supported by partly distinct but interacting systems [DST14]. This distinction matters for the evaluation of LLM responses because perceived understanding and perceived emotional support may not appear equally strongly in the same response.

2.3 The EPITOME Framework

The EPITOME framework (*EmPathy In Text-based, asynchrOnous MEntal health conversations*) was introduced by Sharma et al. to analyse empathy in written, asynchronous support conversations [SMAA20]. This setting is relevant to the present study because empathy is evaluated through written responses only. In face-to-face interaction, empathy can also be communicated through voice, facial expression, and body language. These cues are absent in written interaction, so EPITOME focuses on how empathy is expressed through language.

In EPITOME, expressed empathy is described through three communication mechanisms: *Emotional Reactions*, *Interpretations*, and *Explorations*[SMAA20]. *Emotional Reactions* refer to the emotional tone of a response, including expressions of warmth, compassion, concern, or care toward the person sharing the experience. This mechanism is closest to affective empathy, because it captures whether the response sounds emotionally supportive. Sharma et al. distinguish weaker and stronger forms of emotional reaction. A weaker response may offer general reassurance, such as *"Everything will be fine"*, while a stronger response directly expresses emotion toward the seeker, such as *"I feel really sad for you"*[SMAA20].

The second mechanism, *Interpretations*, concerns the understanding communicated in a response. A response shows interpretation when it reflects what the other person may be feeling or going through [SMAA20]. This mechanism is especially relevant to cognitive empathy, because it concerns whether the response identifies and communicates an understanding of the person's emotional state. For example, a general sentence such as *"I understand how you feel"* gives only a broad acknowledgement. A more specific interpretation, such as *"This must be terrifying"*, shows a clearer inference about the person's experience [SMAA20].

The third mechanism, *Explorations*, captures attempts to understand the seeker further by asking about feelings or experiences that are not yet fully clear [SMAA20]. A broad question such as *"What happened?"* is weaker than a question that refers more directly to the seeker's possible emotional experience [SMAA20]. Exploration is most relevant in longer conversations, where the supporter can ask follow-up questions and build understanding over multiple turns. Since the present study evaluates single-turn responses, Emotional Reactions and Interpretations are the most central mechanisms for the rating procedure.

For each mechanism, EPITOME distinguishes between no communication, weak communication, and strong communication of empathy. Sharma et al. applied the framework to approximately 10,000 post-response pairs from online mental health support platforms, using trained annotators to label the empathy mechanisms in the responses [SMAA20]. This makes EPITOME useful for evaluating empathy as it appears in written responses.

Because the present study evaluates single-turn written responses, EPITOME is used mainly through its Emotional Reactions and Interpretations mechanisms. Emotional Reactions inform the affective empathy rating, while Interpretations inform the cognitive empathy rating. The Method section explains how these mechanisms were adapted into the final scoring procedure.

2.4 LLMs and Empathy

Empathy in LLM responses should be understood as something communicated through generated text, not as evidence that the model has an internal emotional experience. This distinction is important because an LLM can produce language that appears emotionally aware or supportive

while still operating as a text-generation system. Sorin et al. reviewed early studies on LLMs and empathy and found that the included work mainly examined ChatGPT-3.5, often in healthcare-related settings, and usually relied on subjective human assessment rather than objective empathy metrics [SBB+24]. The review also reports that LLMs were evaluated on empathy-related behaviours such as emotion recognition and emotionally supportive response generation [SBB+24]. At the same time, Sorin et al. note several difficulties in interpreting these evaluations, including the possibility that longer responses may be judged as more empathic and the absence of human communication cues such as eye contact and tone of voice [SBB+24]. Lee et al. also treat empathy as something judged by readers, showing that LLM-generated responses can be perceived as empathic by human evaluators [LSZ+24]. Thus empathy in LLM-generated text is best evaluated based on the extent the response appears to understand the user's emotional situation and how supportively it responds to that situation.

2.5 Testing Empathy in LLMs Empirically

Previous research on LLM empathy has used different evaluation methods. Some studies examine how human readers judge generated responses, while others use psychological questionnaires to compare LLMs with human empathy measures. These approaches focus on different aspects of the problem. Human-rating studies evaluate perceived empathy in written responses, whereas questionnaire-based studies examine whether model behaviour follows cognitive and affective empathy dimensions used in human psychology.

Lee et al. studied perceived empathy by asking human raters to evaluate LLM-generated responses to posts about everyday emotional experiences, including workplace situations, parenting, relationships, anxiety, and anger [LSZ+24]. Their results showed that LLM-generated responses were rated as more empathic than human-written responses. At the same time, the authors noted that model outputs sometimes contained recognizable language patterns, such as repeated punctuation, emoji use, and similar wording across responses [LSZ+24]. This is relevant for response-based evaluation because a generated reply may appear empathic while still showing signs of formulaic language.

Scenario-based work has also examined empathy through separate response dimensions. Liu et al. asked both LLMs and human participants to respond to empathy-related scenarios, after which another group of participants rated the responses on a 0–10 scale [LCHW25]. Their rating method was based on EPITOME and focused on two dimensions: Emotional Reactions and Interpretations. Emotional Reactions were used to represent emotional empathy, while Interpretations were used to represent cognitive empathy [LCHW25]. This approach is close to the design of the present study because it evaluates empathy in responses to scenarios and separates emotional support from emotional understanding.

Other studies have used psychological questionnaires rather than ratings of generated replies. Yu et al. evaluated GPT-4 and Llama3 using the Interpersonal Reactivity Index and the Basic Empathy Scale [YPF+25]. Their results showed that GPT-4 reproduced the cognitive and affective empathy structure found in humans, but still showed lower empathy abilities than human participants. Llama3 did not show the same human-aligned empathy structure [YPF+25]. Williams and Rosman also compared cognitive and affective empathy in language models using standardized psychological tests. They found that LLMs performed strongly on cognitive empathy tasks, while both small and large language models showed lower affective empathy than human participants [WR25]. The findings from Liu et al. and Williams and Rosman should be interpreted with caution, as both

studies are preprints.
The literature supports two choices made in this study which are evaluating empathy in generated responses and separating cognitive from affective empathy. What remains less clear is whether empathy ratings change when the same emotional situation is presented implicitly or with an explicit emotion label. Existing scenario-based studies have evaluated empathic responses, and questionnaire-based studies have compared empathy dimensions, but the specific contrast between implicit and explicit emotional context has received less attention. The present study addresses this gap by comparing implicit and explicit versions of the same emotional scenarios while evaluating cognitive and affective empathy separately.

2.6 Research Relevance

The present study is motivated by a gap in how emotional context is treated in LLM empathy evaluation. Previous work has shown that LLM-generated responses can be perceived as empathic by human evaluators, and that cognitive and affective empathy may differ in model behaviour [LSZ+24, LCHW25, YPF+25]. However, less attention has been given to how the user's emotion is expressed in the prompt itself. While the explicit–implicit emotion distinction has been examined in emotion recognition, where models are evaluated on directly expressed and contextually inferred emotional cues [LLQX26], it has received little attention in the evaluation of perceived empathy in generated LLM responses. In everyday interactions, users may describe difficult or meaningful events without explicitly stating that they feel guilty, anxious, happy, or disappointed. This requires the model to infer the emotional state from contextual cues, whereas a directly named emotion provides a clearer signal that may guide both the interpretation of the user's situation and the emotional tone of the response. To examine this factor, the study compares implicit and explicit versions of the same emotional scenarios. The same underlying situation is used across conditions, which allows the comparison to focus on emotional explicitness rather than differences in scenario content. The separate evaluation of cognitive and affective empathy further allows the study to examine whether explicit emotion labels influence perceived emotional understanding and perceived emotional support in similar or different ways.

2.7 Selected LLMs for Empathy Evaluation

This study evaluates two instruction-tuned models from the OLMo model line: OLMo-2 7B Instruct and OLMo-3 7B Instruct. The models were selected because they come from the same open model project, are available at the same 7B scale, and are both instruction-tuned models. Using models from the same model line and at the same parameter scale helps keep the comparison more controlled than a comparison between unrelated models.
OLMo-2 is a family of dense autoregressive language models released at 7B, 13B, and 32B parameter scales. The OLMo-2 release includes model weights, training data, training code, training recipes, training logs, and intermediate checkpoints [OLM25]. OLMo-2 7B Instruct is the instruction-tuned 7B model used in this experiment. The OLMo-2 Instruct models were developed through post-training, including supervised finetuning, preference tuning, and reinforcement learning with verifiable rewards [OLM25]. The OLMo-2 report evaluates these instruction-tuned models on tasks such as instruction following, math, knowledge reasoning, and safety [OLM25]. For the present

study, OLMo-2 7B Instruct was suitable because the experiment required the model to respond to user-written emotional scenarios in a conversational format.
OLMo-3 is the newer model family in this comparison. It is described as a fully open model family at the 7B and 32B scales, with access to the full model development flow, including training stages, checkpoints, data, and dependencies [OLM26]. The OLMo-3 model family targets capabilities such as long-context reasoning, function calling, coding, instruction following, general chat, and knowledge recall [OLM26]. The version used in this study is OLMo-3 7B Instruct. This variant is trained to produce shorter and more direct responses to user queries without generating intermediate thinking traces [OLM26]. This fits the present experiment, because the study evaluates short single-turn responses rather than extended reasoning traces.
The comparison between OLMo-2 and OLMo-3 is not intended to represent all language models. Instead, it compares two successive instruction-tuned OLMo models under the same experimental conditions. Keeping the model line, parameter scale, prompt format, and generation setup consistent helps reduce unnecessary differences between the model conditions. This allows the study to focus on whether the two OLMo versions differ in perceived cognitive and affective empathy. For readability, the remainder of this thesis refers to OLMo-2 7B Instruct as OLMo-2 and OLMo-3 7B Instruct as OLMo-3.

3 Research Questions

The central research question of this thesis is:

> *How does the level of emotional context explicitness in user scenarios influence perceived cognitive and affective empathy in responses generated by OLMo-2 and OLMo-3?*

This question focuses on whether explicitly naming a speaker's emotional state changes how empathic a model's response is perceived to be. In the implicit condition, the emotional state is conveyed through the situation but is not directly named. In the explicit condition, the same scenario is used, but the speaker's emotional state is directly labelled. Since empathy is treated as a multidimensional construct, cognitive and affective empathy are evaluated separately.
To answer the main research question, this study focuses on four subquestions:

- Do explicit emotion labels affect perceived cognitive and affective empathy scores?
- Do OLMo-2 and OLMo-3 differ in perceived cognitive and affective empathy scores?
- Does the effect of emotional explicitness differ between OLMo-2 and OLMo-3?
- Are perceived cognitive and affective empathy scores different within each model?

These research questions are answered through the statistical analysis described in the Method section. The first two subquestions are tested through the main effects of emotional explicitness and model version in the repeated-measures ANOVAs. The third subquestion is tested through the model-by-explicitness interaction. The fourth subquestion is examined with paired-samples t-tests comparing cognitive and affective empathy scores within each model.

4 Method

4.1 Experimental Setup

This study used a 2×2 within-scenario experimental design to examine how model version and emotional context explicitness influence the perceived empathy of large language model responses. The first independent variable was model version, comparing OLMo-2 7B Instruct and OLMo-3 7B Instruct. The second independent variable was emotional explicitness, comparing implicit and explicit versions of the same emotional scenario. The dependent variables were perceived cognitive empathy and perceived affective empathy, both rated by human evaluators.

This experiment used 24 emotional scenarios adapted from the EmpatheticDialogues dataset, which was developed for research on empathetic open-domain dialogue [RSLB19]. Each scenario was prepared in an implicit and an explicit version and was then used as input for both models, resulting in 96 generated responses in total for all conditions. Each of the models was prompted to generate a response to the scenario shared by a person in no more than 150 words. The responses were systematically evaluated by the selected human evaluators using an adapted version of the EPITOME framework. EPITOME was originally developed to characterize expressed empathy in text-based, asynchronous conversations and distinguishes between Emotional Reactions, Interpretations, and Explorations [SMAA20]. In this study, the framework was adapted to evaluate two dimensions: cognitive empathy, referring to understanding of the person's situation or emotional state, and affective empathy, referring to warmth, care, compassion, or emotional support. The detailed training of the evaluation framework and scoring procedure are described in Section 4.4.

The scenario was treated as the repeated-measures unit in the statistical analysis. Each base scenario was compared across four experimental conditions: OLMo-2 implicit, OLMo-2 explicit, OLMo-3 implicit, and OLMo-3 explicit versions. The number of base scenarios needed for this experiment was suitable for future repeated-measures analysis. An a priori power analysis was conducted in G*Power, a power-analysis program commonly used in behavioral and social science research [FELB07, FEBL09]. The power analysis assumed a medium effect size of $f = 0.25$, $\alpha = .05$ desired power of .80, one group, four repeated measurements, a correlation of .50 among repeated measures, and a nonsphericity correction of $\epsilon = 1$. Under these conditions, 24 scenario-level observations provided approximately .80 power, with an actual power of .817, while keeping the rating task manageable and not very exhausting for human evaluators.

After response generation, the outputs were distributed across four Qualtrics survey blocks using a Latin Square counterbalancing design. Each block contained one version of each base scenario, so that no evaluator rated multiple versions of the same scenario. Each generated response was rated by four independent evaluators overall. The following sections describe the scenario materials, response generation procedure, survey design, data cleaning, and statistical analysis in more detail.

Also, the full project repository is available on the author's GitHub page: https://github.com/melisauzun4/Bachelor_Thesis_LLM_Empathy. This repository includes the scenario materials, response-generation materials, cleaned and processed data files, the code of data analysis, statistical outputs, plots and other supplementary materials.

4.2 Scenario Materials and Conditions

The scenario materials were adapted from the EmpatheticDialogues dataset introduced by Rashkin et al. [RSLB19]. This dataset was developed as a benchmark for empathetic open-domain dialogue generation. Its purpose was to support models in recognizing a conversation partner's emotional situation and responding in an appropriate and empathetic way. The dataset contains approximately 25,000 crowdsourced one-on-one conversations grounded in emotional situations. In the original task, one worker acted as the speaker and was given an emotion label. The speaker then wrote a short personal situation associated with that emotion and discussed it with another worker acting as the listener. The listener did not see the original emotion label or situation description, and therefore had to respond based on the emotional cues expressed in the conversation. This makes the dataset suitable for the present study, because it contains short personal situations associated with emotion labels while also reflecting the problem of responding to emotions that may need to be inferred from context.

From the EmpatheticDialogues dataset, 24 base scenarios were selected and adapted into short first-person prompts. The final scenario set included both positive and negative emotional situations, with 12 positive and 12 negative base scenarios, using the emotion labels from the EmpatheticDialogues dataset. This was done to avoid limiting the study to only distress-related situations and to include a broader range of emotional contexts. Valence was used to balance the materials, but it was not included as a main factor in the primary statistical analysis.

Each base scenario was rewritten into an implicit and an explicit version. The implicit version still contained emotional cues through the described situation, but it did not directly name the speaker's emotional state. In other words, the model had to infer the emotion from the content of the personal experience. The explicit version preserved the same situational content but added a short final sentence that directly stated the emotion. This added sentence was based on the emotion label associated with the original EmpatheticDialogues scenario and was phrased in a natural first-person form, such as "I felt really guilty" or "I feel really happy", depending on the scenario.

For example, in the evaluator training material, the implicit scenario described a person seeing a turtle on the road, being unable to stop, and hoping that the turtle survived. This version contained cues of guilt and concern, but did not directly state the emotion. The explicit version kept the same situation and added the sentence "I felt really guilty." This manipulation allowed the underlying situation to remain constant while varying whether the emotional state was only implied or directly available to the model.

The aim of this design was to isolate the effect of emotional explicitness as much as possible. Therefore, the implicit and explicit versions of each scenario were kept close in content and style, differing mainly in the presence or absence of the explicit emotion sentence. The scenarios were written as personal experiences shared by a person, so that the models were required to respond to an emotionally meaningful disclosure rather than to answer a factual question. The exact response generation procedure is described in the next section.

4.3 Response Generation

Responses were generated locally using Ollama. Before generation, both models were pulled through Ollama using their exact model identifiers. The OLMo-2 model used in this study was pulled with

the identifier `olmo2:7b-1124-instruct-q8_0`, and the OLMo-3 model was pulled with the identifier `olmo-3:7b-instruct-q8_0`. These identifiers correspond to 8-bit quantized Ollama variants of OLMo-2 7B Instruct and OLMo-3 7B Instruct, respectively [Olld, Ollc]. The OLMo-2 and OLMo-3 model families, including their instruct variants, are described in the original OLMo technical reports [OLM25, OLM26]. Running the models locally made the response generation procedure independent of external API availability and allowed the same setup to be applied consistently across both models.

The responses were generated with Python scripts that sent each scenario to Ollama's local chat endpoint, `localhost:11434/api/chat`. Ollama's API supports chat requests with a list of messages and allows generation options to be specified in the request body [Olla]. Each scenario was sent independently as a single user message, without previous conversation history. This ensured that no earlier scenario could influence the generation of a later response.

The same prompt format was used for all scenarios and both models:

> *Respond to the following situation shared with you by a person in no more than 150 words. Situation:*

This instruction was placed before the scenario text. The prompt did not explicitly instruct the models to respond empathically because the study aimed to evaluate how empathic the generated responses were under implicit and explicit emotional-context conditions, rather than asking the models to respond empathically directly.

All generation parameters were kept fixed across both models and all scenarios. The scripts used `temperature = 0.0`, `top_p = 1.0`, `seed = 42`, and `num_predict = 250`. According to the Ollama documentation, `temperature` controls the creativity of model outputs, `seed` sets the random seed for generation, `top_p` affects token sampling, and `num_predict` specifies the maximum number of tokens generated [Ollb]. Temperature was set to 0.0 to minimize random variation during response generation, and the seed was fixed at 42 as an additional reproducibility control. The value `num_predict = 250` was selected to allow complete short responses while keeping output length bounded and consistent with the prompt instruction of no more than 150 words. Keeping the same parameter values for both models reduced the chance that differences in generated responses were caused by differences in generation settings.

The API request used `stream = false` so that Ollama returned each response only after it had been fully generated [Olla]. For each response, the script saved the scenario identifier, condition, emotion label, full prompt text, model name, generated response, and generation duration. The responses generated from both models were later used to construct the Qualtrics survey blocks.

4.4 Survey Design and Data Collection

The generated responses were evaluated through a Qualtrics survey. 16 evaluators were recruited from the researcher's academic network. Participants were eligible to participate if they were able to evaluate English text and had enough time to complete the survey in one sitting. The final evaluator group consisted of eight evaluators with a technical or computer science related background and eight evaluators without such a background.

The rating dimensions were adapted from the EPITOME framework [SMAA20]. EPITOME distinguishes three mechanisms of expressed empathy in text-based conversations: Emotional Reactions,

Interpretations, and Explorations. A similar adaptation has been used in recent work on LLM empathy evaluation, where Emotional Reactions and Interpretations were used to assess emotional and cognitive empathy in generated responses [LCHW25]. In the present study, Interpretations were used as the basis for cognitive empathy, because this mechanism concerns communicating an understanding of another person's feelings and experiences. Emotional Reactions were used as the basis for affective empathy, because this mechanism concerns expressing warmth, compassion, concern, or emotional support. Explorations were not included in the rating task because this study evaluated single-turn model responses rather than multi-turn conversations in which follow-up questions could naturally occur.

Evaluators received an EPITOME-based training guide on the second page of the survey. The guide explained that evaluators would read a short scenario followed by a response generated by a LLM, and that each response had to be rated on two independent empathy dimensions. Cognitive empathy was defined as the extent to which the response showed understanding of the person's personal experience and emotional state. Affective empathy was defined as the degree to which the response expressed warmth, care, compassion, or emotional support toward the person.

Both dimensions were rated separately on a 0–10 scale, where 0 represented the least empathic and 10 represented the most empathic. To support consistent use of the scale, the guide divided the ratings into three qualitative bands: 0–3 for low or no empathy, 4–6 for medium empathy, and 7–10 for strong empathy. Evaluators were instructed to first identify the qualitative band that best matched the response and then choose a specific score within that band.

The training guide also included several scoring instructions. Evaluators were asked to rate each response holistically, based on the full response rather than on one phrase in isolation. They were also told that empathy is not the same as helpfulness: advice or factual information alone should not be counted as empathy unless the response also showed emotional understanding and care. Finally, the guide emphasized that cognitive and affective empathy should be judged independently, because a response could show understanding without much warmth, or warmth without much specific understanding.

To make the rating criteria more concrete, the training guide included worked examples. These examples showed how responses to the same scenario could receive different cognitive and affective empathy scores. For example, a response that mainly gave corrective advice was presented as low on both dimensions, while a response that recognized the speaker's guilt and expressed emotional warmth was presented as higher in both cognitive and affective empathy. The full training guide is included in Appendix A.

The 96 generated responses were anonymously distributed across four Qualtrics survey blocks using a Latin Square counterbalancing design. Each block contained 24 responses, with one version of each base scenario. This ensured that no evaluator saw more than one version of the same base scenario, reducing possible carryover effects between implicit and explicit versions or between model responses to the same underlying situation. The four experimental conditions were coded as follows: A = OLMo-2 implicit, B = OLMo-2 explicit, C = OLMo-3 implicit, and D = OLMo-3 explicit. The same rotation continued across all 24 scenarios, and within each Qualtrics block the order of responses was randomized to reduce order effects. Table 1 summarizes the Latin Square rotation pattern.

The table illustrates the rotation pattern using the first four scenarios. For example, in Block 1 the first four scenarios followed the sequence S1-A, S2-B, S3-C, and S4-D, whereas in Block 2 the sequence shifted to S1-B, S2-C, S3-D, and S4-A. This structure ensured that, across the complete

Table 1: Latin Square survey block design. A = OLMo-2 implicit, B = OLMo-2 explicit, C = OLMo-3 implicit, and D = OLMo-3 explicit. Each block contained 24 responses and was completed by four evaluators.

Block	Evaluators	S1	S2	S3	S4
1	R1–R4	A	B	C	D
2	R5–R8	B	C	D	A
3	R9–R12	C	D	A	B
4	R13–R16	D	A	B	C

design, every base scenario was represented in all four model-by-explicitness conditions, while each individual evaluator rated only one version of each base scenario.

Each evaluator rated 24 responses. For every item, evaluators saw the scenario text and the generated model response, but they were not shown the model name or condition label. This was done to reduce expectation bias toward either model or condition. Evaluators then assigned two separate scores: one for cognitive empathy and one for affective empathy. Across the full design, each generated response received four independent ratings per dimension.

Survey links were distributed individually when evaluators were ready to complete the task. Evaluators were asked to complete the survey in one sitting without taking breaks, to reduce interruptions and incomplete submissions. The average completion time for the completed surveys was approximately 32.1 minutes.

4.5 Data Cleaning and Preparation

After data collection, the Qualtrics exports were cleaned and prepared for statistical analysis in Python. The final cleaned dataset contained 16 complete evaluator scores across the four survey blocks.

A correction was applied to the exported slider-scale values. In the Qualtrics survey, evaluators rated cognitive and affective empathy on a visible 0–10 scale. However, the exported values were encoded from 1 to 11. To recover the intended 0–10 scale, all exported empathy score values were uniformly recoded by subtracting 1 from each value. For example, an exported value of 5 was recoded as 4, and an exported value of 11 was recoded as 10. This correction was verified by comparing the recoded values with the scores shown in the Qualtrics results dashboard.

The score columns were then renamed from Qualtrics question codes to descriptive variable names. The new names followed the structure `S[n]_[Model]_[Condition]_[Dimension]`, for example `S1_Olmo2_Implicit_Cognitive`. This made it possible to identify the scenario number, model, explicitness condition, and empathy dimension for each score.

The data were then reshaped from wide format to long format. In the long-format dataset, each row represented one evaluator's score for one scenario, one model, one explicitness condition, and one empathy dimension. Each observation was tagged with the survey block, evaluator ID, scenario ID, model, explicitness condition, and empathy dimension. The resulting long-format dataset contained 768 observations, corresponding to 16 evaluators, 24 scenarios, and two empathy dimensions.

Two analysis datasets were created. The first dataset retained individual evaluator scores and was used for the inter-rater reliability analysis. The second dataset was created by averaging the four evaluator ratings for each generated response. This produced one averaged cognitive empathy score

and one averaged affective empathy score for each response. These averaged response-level scores were used for the repeated-measures ANOVA and paired-samples t-tests.

4.6 Statistical Analysis

All statistical analyses were conducted in Python using Google Colab. The main packages used were `pandas` for data cleaning and reshaping, `pingouin` for inter-rater reliability, repeated-measures ANOVA, and paired-samples t-tests, and `matplotlib` for visualisation. The analysis was based on two prepared datasets. One dataset retained the individual evaluator ratings and was used for the inter-rater reliability analysis. The other dataset contained response-level scores, created by averaging the four evaluator ratings for each generated response, and was used for the repeated-measures ANOVA and paired-samples t-tests.

Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC). Specifically, ICC(A,k), an absolute-agreement, average-measures ICC, was calculated separately for each survey block and each empathy dimension. This type of ICC was selected because the final response-level scores were based on the average of four evaluator ratings, and because agreement on the absolute 0–10 scores was more relevant than only consistency in the ranking of responses. The interpretation of ICC values followed the guideline proposed by Koo and Li, where values below .50 indicate poor reliability, values between .50 and .75 indicate moderate reliability, values between .75 and .90 indicate good reliability, and values above .90 indicate excellent reliability [KL16].

Before the main inferential analyses, the four evaluator ratings for each generated response were averaged. The scenario was then used as the repeated-measures subject unit. This meant that each scenario had one averaged score in each of the four model-by-explicitness conditions: OLMo-2 implicit, OLMo-2 explicit, OLMo-3 implicit, and OLMo-3 explicit. Separate analyses were conducted for cognitive empathy and affective empathy.

Descriptive statistics were first computed for each combination of empathy dimension, model, and explicitness condition. Means and standard deviations were used to inspect the direction of the effects before running inferential tests. The assumptions for repeated-measures ANOVA were then checked. Normality of the relevant difference scores was assessed using the Shapiro-Wilk test. Sphericity was not a concern because each within-subject factor had only two levels: model had the levels OLMo-2 and OLMo-3, and explicitness had the levels implicit and explicit. With two-level within-subject factors, there is only one difference score per factor, so the sphericity assumption is automatically satisfied.

The main inferential analysis consisted of two separate 2×2 repeated-measures ANOVAs: one for cognitive empathy and one for affective empathy. In both analyses, the within-subject factors were model and emotional explicitness. These ANOVAs tested three effects: the main effect of model, indicating whether OLMo-2 and OLMo-3 differed overall; the main effect of explicitness, indicating whether explicit scenarios received different empathy scores than implicit scenarios; and the model-by-explicitness interaction, indicating whether the effect of emotional explicitness differed between OLMo-2 and OLMo-3.

Finally, paired-samples t-tests were conducted to compare cognitive and affective empathy scores within each model. These tests examined whether each model showed a difference between the two empathy dimensions, collapsed across explicitness conditions. One paired-samples t-test compared cognitive and affective empathy for OLMo-2, and a second paired-samples t-test compared cognitive and affective empathy for OLMo-3. Cohen's d was reported as the effect size for these comparisons.

A secondary G*Power calculation was conducted for the paired-samples t-tests. Using a medium effect size of Cohen's $d_z = 0.5$, $\alpha = .05$, and desired power of .80, the required sample size was 34 paired observations. The present paired comparisons used 48 paired observations per model, corresponding to 24 scenarios across two explicitness conditions. This provided sufficient scenario-level observations for the planned within-model comparisons.

5 Experimental Results

5.1 Inter-rater Reliability

Inter-rater reliability was assessed before the main inferential analyses. The ICC results are shown in Table 2. Reliability varied across blocks and empathy dimensions. For cognitive empathy, ICC values ranged from poor to moderate agreement, with values between .162 and .572. For affective empathy, agreement was generally higher, ranging from poor to good agreement, with values between .484 and .780.

Table 2: Inter-rater reliability by survey block and empathy dimension. ICC values were interpreted according to Koo and Li's guidelines: values below .50 indicate poor reliability, values between .50 and .75 indicate moderate reliability, and values between .75 and .90 indicate good reliability [KL16].

Block	Cognitive ICC	Interpretation	Affective ICC	Interpretation
1	.572	Moderate	.780	Good
2	.276	Poor	.484	Poor
3	.361	Poor	.736	Moderate
4	.162	Poor	.519	Moderate

Overall, affective empathy showed more consistent agreement between evaluators than cognitive empathy. This suggests that evaluators found it easier to agree on emotional warmth, care, or support. By contrast, cognitive empathy may have required more interpretation, because raters had to judge whether a response genuinely understood the speaker's situation or only acknowledged it at a surface level.

5.2 Descriptive Statistics

Descriptive statistics were calculated for each combination of model, emotional explicitness condition, and empathy dimension. The means and standard deviations are shown in Table 3. Overall, OLMo-3 received higher empathy scores than OLMo-2 across both dimensions and both explicitness conditions. In addition, explicit scenarios generally received higher empathy scores than implicit scenarios. For cognitive empathy, OLMo-2 increased from a mean score of 5.45 in the implicit condition to 6.31 in the explicit condition. OLMo-3 showed a similar increase, from 6.38 in the implicit condition to 7.31 in the explicit condition. This indicates that explicit emotional context was associated with higher perceived cognitive empathy for both models. For affective empathy, the same general pattern was observed. OLMo-2 increased from 4.17 in the implicit condition to 5.38 in the explicit condition, while OLMo-3 increased from 6.26 to 7.22. The difference between

Table 3: Descriptive statistics for cognitive and affective empathy scores by model and emotional explicitness condition.

Dimension	Condition	Mean	SD
Cognitive	OLMo-2 Implicit	5.45	1.28
Cognitive	OLMo-2 Explicit	6.31	1.19
Cognitive	OLMo-3 Implicit	6.38	1.01
Cognitive	OLMo-3 Explicit	7.31	1.03
Affective	OLMo-2 Implicit	4.17	1.33
Affective	OLMo-2 Explicit	5.38	1.37
Affective	OLMo-3 Implicit	6.26	1.29
Affective	OLMo-3 Explicit	7.22	1.37

the models was especially visible for affective empathy: OLMo-3 scored substantially higher than OLMo-2 in both the implicit and explicit conditions.

Overall the descriptive statistics suggest two main patterns. First, explicit scenarios had higher perceived empathy scores than implicit scenarios. Second, OLMo-3 received higher empathy ratings than OLMo-2 across all conditions. These patterns were tested statistically in the repeated-measures ANOVAs reported in the next section.

5.3 Repeated-Measures ANOVA

Before conducting the repeated-measures ANOVAs, the normality of the relevant difference scores was assessed using the Shapiro-Wilk test. For cognitive empathy, the model difference scores were normally distributed, $p = .526$, and the explicitness difference scores also did not significantly deviate from normality, $p = .068$. For affective empathy, the model difference scores were normally distributed, $p = .990$, as were the explicitness difference scores, $p = .369$. Since all Shapiro-Wilk tests were non-significant, the normality assumption was considered sufficiently met. Sphericity was not a concern because both within-subject factors, model and emotional explicitness, had only two levels.

Two separate 2×2 repeated-measures ANOVAs were conducted: one for cognitive empathy and one for affective empathy. In both analyses, the within-subject factors were model, with the levels OLMo-2 and OLMo-3, and emotional explicitness, with the levels implicit and explicit.

For cognitive empathy, the repeated-measures ANOVA showed a significant main effect of model, $F(1, 23) = 18.21$, $p < .001$, $\eta^2_G = .16$. This means that the null hypothesis of no model effect was rejected for cognitive empathy. OLMo-3 received higher cognitive empathy scores than OLMo-2 in both the implicit and explicit conditions. The analysis also showed a significant main effect of emotional explicitness, $F(1, 23) = 28.34$, $p < .001$, $\eta^2_G = .14$. Therefore, the null hypothesis of no explicitness effect was also rejected for cognitive empathy. Explicit scenarios received higher cognitive empathy scores than implicit scenarios. However, the interaction between model and emotional explicitness was not significant, $F(1, 23) = 0.02$, $p = .890$, $\eta^2_G < .001$. Thus, the analysis failed to reject the null hypothesis of no model-by-explicitness interaction for cognitive empathy.

Figure 1 visualizes the cognitive empathy results. The plot shows that both models received higher cognitive empathy scores in the explicit condition than in the implicit condition. It also shows that OLMo-3 was rated higher than OLMo-2 in both conditions. The roughly parallel lines are

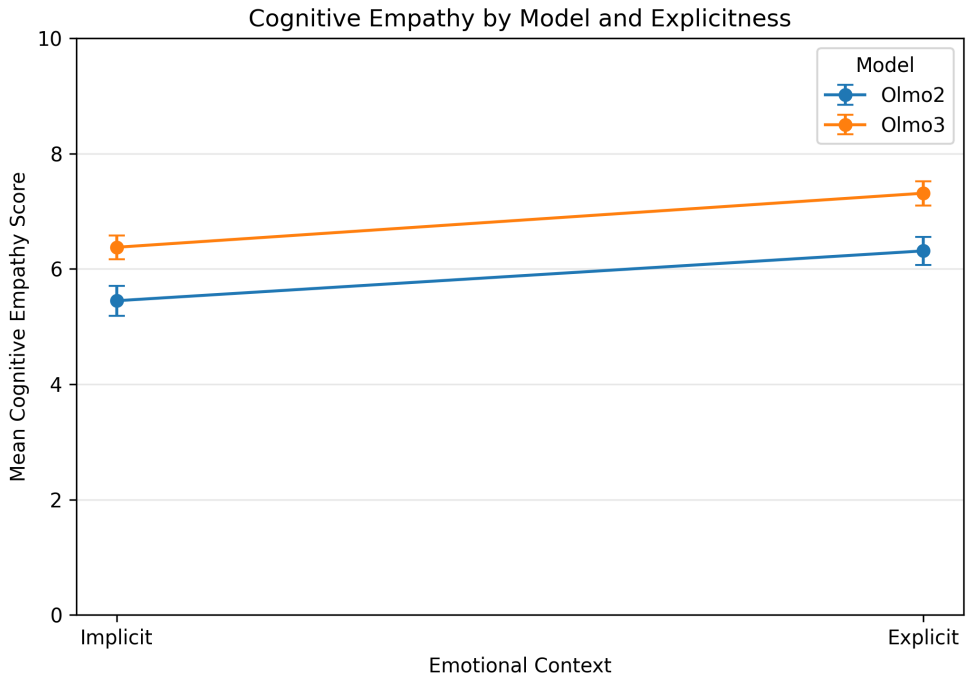

Figure 1: Mean cognitive empathy scores by model and emotional explicitness condition. Error bars indicate standard error.

consistent with the non-significant interaction effect, indicating that the increase from implicit to explicit scenarios was comparable across the two models.

For affective empathy, the repeated-measures ANOVA also showed a significant main effect of model, $F(1, 23) = 78.05$, $p < .001$, $\eta^2_G = .36$. This means that the null hypothesis of no model effect was rejected for affective empathy. OLMo-3 received higher affective empathy scores than OLMo-2. In addition, there was a significant main effect of emotional explicitness, $F(1, 23) = 18.24$, $p < .001$, $\eta^2_G = .15$. Therefore, the null hypothesis of no explicitness effect was also rejected for affective empathy. Explicit scenarios received higher affective empathy scores than implicit scenarios. The interaction between model and emotional explicitness was not significant, $F(1, 23) = 0.18$, $p = .679$, $\eta^2_G = .002$. Thus, the analysis failed to reject the null hypothesis of no model-by-explicitness interaction for affective empathy.

Figure 2 visualizes the affective empathy results. The plot shows the same general pattern as the cognitive empathy analysis: explicit scenarios received higher scores than implicit scenarios, and OLMo-3 received higher scores than OLMo-2 across both explicitness conditions. The distance between the two model lines is larger for affective empathy than for cognitive empathy, which is consistent with the larger generalized eta-squared value for the model effect in the affective empathy ANOVA. However, because the interaction was not significant, the results do not indicate that emotional explicitness affected the two models differently.

Overall, the ANOVA results led to the rejection of the null hypotheses of no model effect and no explicitness effect for both cognitive and affective empathy. Specifically, OLMo-3 received significantly higher empathy scores than OLMo-2, and explicit scenarios received significantly higher empathy scores than implicit scenarios. Therefore, the alternative hypotheses that model

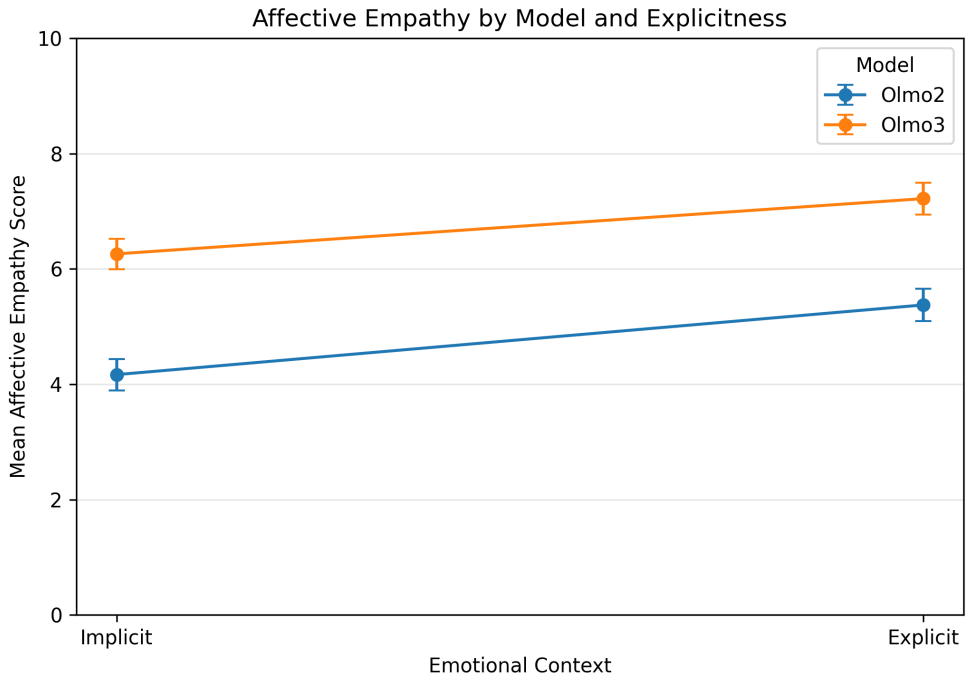

Figure 2: Mean affective empathy scores by model and emotional explicitness condition. Error bars indicate standard error.

version and emotional explicitness influence perceived empathy were supported. However, the analyses failed to reject the null hypothesis of no model-by-explicitness interaction for either cognitive or affective empathy. Thus, the alternative hypothesis that the effect of emotional explicitness would differ between OLMo-2 and OLMo-3 was not supported.

5.4 Paired t-tests

Paired-samples t-tests were conducted to examine whether cognitive and affective empathy scores differed within each model. These tests addressed the question of whether each model showed a difference between the two empathy dimensions when collapsed across emotional explicitness conditions. For each model, the paired observations consisted of the averaged cognitive and affective empathy scores for the same scenario-condition combinations. This resulted in 48 paired observations per model, corresponding to 24 scenarios across the two explicitness conditions.

For OLMo-2, cognitive empathy scores were significantly higher than affective empathy scores, $t(47) = 11.17$, $p < .001$, Cohen's $d = 0.80$. Therefore, the null hypothesis of no difference between cognitive and affective empathy was rejected for OLMo-2. This supports the alternative hypothesis that cognitive and affective empathy scores differ within OLMo-2.

For OLMo-3, the difference between cognitive and affective empathy scores was not significant, $t(47) = 0.75$, $p = .458$, Cohen's $d = 0.08$. Therefore, the analysis failed to reject the null hypothesis of no difference between cognitive and affective empathy for OLMo-3. This means that the alternative hypothesis that cognitive and affective empathy scores differ within OLMo-3 was not supported.

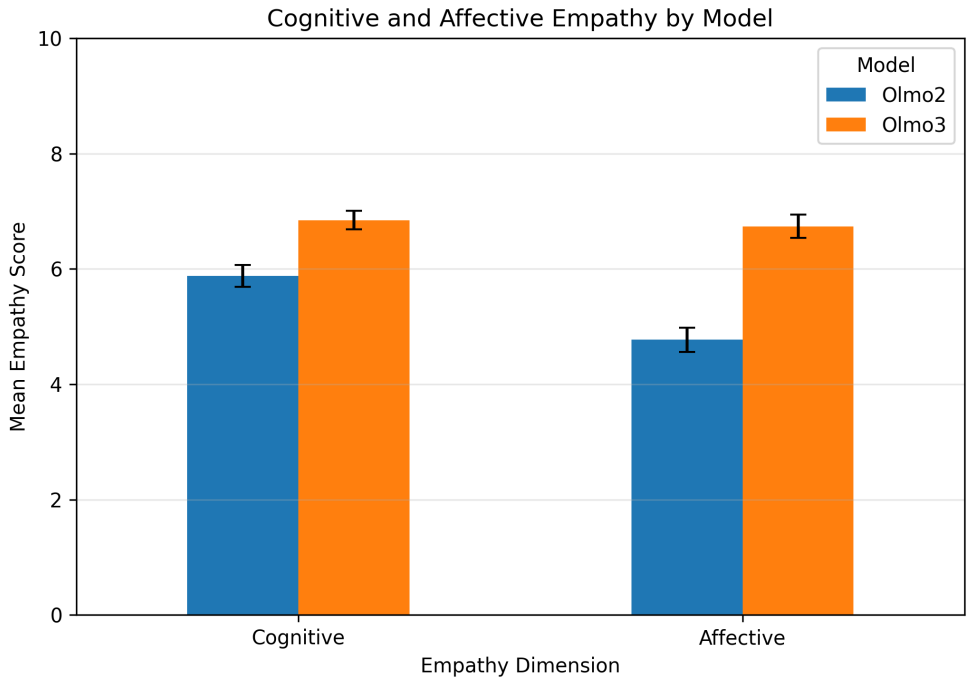

Figure 3: Mean cognitive and affective empathy scores by model, collapsed across emotional explicitness conditions. Error bars indicate standard error.

Figure 3 visualizes the mean cognitive and affective empathy scores for both models. The figure shows a clear separation between cognitive and affective empathy for OLMo-2, with cognitive empathy receiving higher scores. In contrast, the two empathy dimensions are much closer for OLMo-3, which is consistent with the non-significant paired t-test result. Overall, these findings suggest that OLMo-2 showed an imbalance between cognitive and affective empathy, whereas OLMo-3 showed a more balanced pattern across the two dimensions.

6 Discussion

This thesis investigated whether emotional context explicitness influences perceived cognitive and affective empathy in responses generated by OLMo-2 and OLMo-3. The results suggest that emotional explicitness affected how empathic the generated responses were perceived to be. Across the scenario set, responses to explicit prompts received higher empathy ratings than responses to implicit prompts. This should be interpreted as an average effect, not as evidence that every explicit scenario produced a more empathic response in every individual case.

One explanation is that explicit emotion labels reduced ambiguity in the prompt. In the implicit condition, the model had to infer the user's emotional state from situational cues. In the explicit condition, the emotion was directly available in the final sentence of the prompt. This may have made it easier for the model to identify the relevant emotional state and to produce a response that validated or supported that emotion.

However, this interpretation should be qualified. Higher empathy ratings in the explicit condition may not only reflect deeper emotional understanding. They may also partly reflect label echoing:

when the prompt states an emotion directly, such as *"I feel really lonely"* or *"I feel very hopeful,"* the model can reuse that label without needing to infer it from the situation. This is especially relevant for cognitive empathy, because a response that repeats the named emotion may appear to understand the user's state even if the model has not produced a more elaborate interpretation of the situation.

To examine this possibility, a simple post-hoc strict label-reuse check was conducted. Each response was coded as 1 if the exact target emotion word appeared at least once, and 0 otherwise. This binary coding was used because the aim was to test whether the model reused the emotion label at all, rather than how many times the same label was repeated within a response. Variants and synonyms were not counted, because words such as *hope* or *care* often appear as part of ordinary supportive language rather than as a direct repetition of the target emotion label. Counting them as reuse would risk mistaking general supportive wording for genuine label echoing, so focusing on exact target emotion words kept the check a conservative estimate of direct label reuse.

Table 4: Strict target emotion-label reuse by emotional explicitness condition.

Condition	**Exact label reused**	**Total responses**	**Percentage**
Implicit	5	48	10.4%
Explicit	23	48	47.9%

As shown in Table 4, exact target-label reuse was more common in the explicit condition than in the implicit condition. This suggests that the explicitness effect may partly reflect lexical cueing: when the emotion label was directly available in the prompt, the models sometimes reused that exact label in the response. Descriptively, the same pattern appeared within both models, with exact label reuse being more frequent in explicit than implicit responses for both OLMo-2 and OLMo-3, although OLMo-3 reused exact labels more often overall.

At the same time, label reuse does not fully explain the empathy ratings. The check captures only one surface feature of the response: whether the exact emotion word appeared at least once. It does not capture whether the response was warm, validating, specific, or advice-oriented. Two responses can both reuse an emotion label but still differ substantially in perceived empathy depending on whether they elaborate on the user's situation and communicate emotional support. The explicitness effect should therefore be understood as involving both reduced emotional ambiguity and a possible label-echoing effect. The finding still shows that prompt wording shapes perceived empathy in generated text, but it should not be interpreted as direct evidence that explicit prompts produce deeper empathic understanding.

The model comparison provides a second interpretation of the findings. OLMo-3 responses were rated as more empathic than OLMo-2 responses across both cognitive and affective empathy. This indicates that, under the same experimental conditions, OLMo-3 was more often perceived as understanding the user's situation and responding in an emotionally supportive way. The difference was especially clear for affective empathy. This suggests that the main distinction between the two models was not only whether they recognized the emotional situation, but also how much care, validation, and emotional support their responses communicated.

This model difference can be related to the model descriptions, but only cautiously. OLMo-2 and OLMo-3 were selected from the same OLMo model line and compared at the same 7B scale. OLMo-2 Instruct is described as an instruction-tuned model developed through post-training,

including supervised finetuning, preference tuning, and reinforcement learning with verifiable rewards [OLM25]. OLMo-3 is a newer model generation, and OLMo-3 Instruct is described as being tuned for shorter and more direct responses without intermediate thinking traces, with a focus on general chat and function calling [OLM26]. This description is consistent with the type of short, direct conversational replies evaluated in this study. However, the experiment did not isolate training data, post-training data, preference tuning, reinforcement learning, or response-style optimization as separate causal factors. The results therefore show that OLMo-3 responses were perceived as more empathic in this setting, but they cannot determine which specific part of the model development caused this difference.

The non-significant interaction between model version and emotional explicitness further clarifies the pattern. Emotional explicitness improved perceived empathy ratings in a broadly similar way for both OLMo-2 and OLMo-3. The effect of explicitly naming the user's emotion was therefore not specific to one model version. This suggests that emotional explicitness acted as a prompt-level factor in this experiment: both models received higher ratings when the emotional state was directly stated. At the same time, this does not mean that the two models responded to explicit emotion labels in exactly the same way. Individual responses still differed in wording, style, and emotional tone. The statistical result only indicates that the size of the explicitness effect was not significantly different between the two models across the full scenario set.

The within-model comparison further clarifies the difference between OLMo-2 and OLMo-3. For OLMo-2, cognitive empathy scores were significantly higher than affective empathy scores. This suggests that OLMo-2 was better at producing responses that appeared to understand or address the user's situation than responses that sustained emotional warmth. The important point is not only that the two scores differed, but that they point to a difference in response style. OLMo-2 often recognized the relevant situation or emotional problem, but then moved quickly toward advice, resources, or practical next steps. This advice-oriented pattern may make the response seem cognitively appropriate, because it shows that the model has identified what kind of problem the user is facing. However, it can also make the response feel less emotionally supportive when validation, warmth, and emotional presence remain brief or secondary.
This pattern is especially relevant because it fits the pattern described in Williams and Rosman's *Heartificial Intelligence* preprint, where language models are suggested to perform better on cognitive aspects of empathy than on affective empathy [WR25]. In the present study, this pattern was specifically visible for OLMo-2: this does not mean that the model was unempathic overall, but that its empathy was more strongly expressed through problem recognition and practical guidance than through sustained emotional support.
For OLMo-3, cognitive and affective empathy scores did not differ significantly. This suggests a more balanced profile: OLMo-3 was perceived not only as understanding the user's situation, but also as responding with a similar level of emotional support. The difference between the two models therefore should not be interpreted only as OLMo-3 being generally better. Rather, OLMo-3 appeared better at combining situational understanding with emotional validation, while OLMo-2 showed a clearer separation between understanding the problem and emotionally supporting the user. This supports the decision to evaluate cognitive and affective empathy separately, because a single overall empathy score would have hidden the fact that OLMo-2 scored higher on cognitive than affective empathy while OLMo-3 did not show the same imbalance [YPF+25, LCHW25].

The results also show why helpfulness and empathy should not be treated as identical. Some

OLMo-2 responses were useful in a practical sense, but practical usefulness did not always lead to high affective empathy ratings. A response can give relevant suggestions while still feeling emotionally distant if it does not validate the user's feelings or communicate care. This distinction was important in the rating procedure, where advice or information alone was not treated as strongly empathic unless it was combined with emotional understanding and warmth. The stronger affective-empathy ratings for OLMo-3 suggest that its responses more often combined situational relevance with supportive language.

The findings connect to previous work on LLM empathy in two ways. Earlier research has shown that LLM-generated responses can be perceived as empathic by human evaluators [LSZ+24]. Other studies have emphasized that cognitive and affective empathy should be examined separately in LLM evaluation [YPF+25, LCHW25]. The present study adds to this work by focusing on emotional explicitness as a specific prompt-level factor. The results suggest that the way a user expresses an emotion can influence how empathic a model response appears. When emotions are only implied, the model has to infer the emotional meaning from context. When the emotion is directly labelled, the model may produce a more direct emotional reflection. This shows that prompt formulation can shape perceived empathy in LLM-generated responses.

The selected responses illustrate how the statistical patterns appeared in actual model outputs. This was not a separate systematic qualitative coding analysis, but an interpretive use of examples based on the same EPITOME-inspired distinction between *Interpretations* and *Emotional Reactions* used in the rating procedure [SMAA20]. The full selected responses are included in Appendix C. In the negative caregiver scenario, the implicit version described a person who had been taking care of an aging mother for five years and had little time for friends, while the explicit version added the sentence "I feel really lonely." In the implicit condition, OLMo-2 recognized the burden of caregiving by stating that caring for an aging parent can be demanding and can leave little time for social activities or personal interests. However, the response then shifted into a list-like sequence of practical suggestions, including support groups, respite care, small moments of self-care, and online communities. This structure matters because it frames the emotional disclosure mainly as a practical problem to be managed. The suggestions may be useful, but the numbered advice format makes the response feel more procedural and less emotionally present. In the explicit condition, OLMo-2 more directly reflected the named emotion by acknowledging that the user was feeling lonely and by stating that the user's feelings were valid. This shows that explicitness helped OLMo-2 produce clearer emotional acknowledgement. However, even in this version, the response still moved quickly toward resources and coping options, suggesting that this advice-oriented style remained present in this selected response. OLMo-3 handled the same caregiver scenario differently. Its responses also included supportive suggestions, but these were embedded within a more sustained validating frame. In the implicit condition, OLMo-3 emphasized that it was understandable to feel isolated when so much energy was needed elsewhere and ended with the sentence "You deserve connection too." In the explicit condition, it connected the user's loneliness to the emotional toll of caregiving and stated that the user deserved support and kindness. Compared with OLMo-2, OLMo-3 kept the user's emotional burden more central throughout the response, rather than foregrounding a list of solutions. This helps explain why OLMo-3 was rated higher on affective empathy: it did not only identify the situation, but also maintained validation and emotional support while responding to it.

A similar pattern appeared in the positive law-firm scenario. The implicit version described applying for a job at a law firm the speaker had always wanted to work for, while the explicit version

added the sentence "I feel very hopeful." OLMo-2 recognized the importance of the opportunity by describing the application as a significant step toward the user's career goals. However, the response again moved quickly toward practical next steps, such as preparing a resume, practicing interview questions, and reaching out to contacts. This suggests that, in the selected examples, the advice-oriented style was not limited to the distress-related scenario. In this positive example as well, OLMo-2 placed more emphasis on practical next steps than on sustaining the user's emotional state. In the explicit condition, OLMo-2 became warmer by describing the opportunity as exciting and dream-like, but the response still focused strongly on preparation and patience during the application process. By contrast, OLMo-3 placed more emphasis on encouragement and shared positive emotion. In the implicit condition, it described applying to the dream firm as a big step and ended by offering to listen and cheer the user on. In the explicit condition, it directly reflected the user's hope and framed the application as a possible turning point. These examples are consistent with the quantitative pattern in which OLMo-3 received higher perceived empathy ratings overall and OLMo-2 scored higher on cognitive than affective empathy. They also show why emotional explicitness mattered: when the emotion was directly named, both models could reflect it more clearly, but the models still differed in how much they sustained emotional support beyond the label itself.

The inclusion of both positive and negative scenarios is also relevant for interpreting the explicitness effect. Valence was used to balance the scenario materials, but it was not included as an independent factor in the main statistical analyses. Descriptively, however, the explicitness pattern appeared in both valence groups. Collapsed across models and empathy dimensions, negative scenarios increased from a mean score of 5.41 in the implicit condition to 6.43 in the explicit condition, while positive scenarios increased from 5.71 to 6.68. The same descriptive pattern was visible when cognitive and affective empathy were considered separately. For negative scenarios, affective empathy increased from 5.01 to 6.12 and cognitive empathy from 5.81 to 6.74. For positive scenarios, affective empathy increased from 5.42 to 6.47 and cognitive empathy from 6.01 to 6.89. These values suggest that the explicitness effect was not limited to negative or distress-related situations. The positive law-firm example illustrates a subtler version of the same response-style difference: OLMo-2 recognized the importance of the opportunity but moved quickly toward practical preparation, whereas OLMo-3 more strongly sustained encouragement and positive emotional support. Future research could examine this more directly by including valence as an independent variable.

A further strength of the present design is that the models were not explicitly instructed to be empathic. The prompt only asked each model to respond to the situation shared by the user in no more than 150 words. This differs from studies such as Liu et al., where models were explicitly instructed to demonstrate empathy [LCHW25]. In the present study, the observed differences in perceived empathy therefore did not simply result from telling the models to be empathic, but emerged from the models' default response tendencies under the same prompt format. At the same time, the study also highlights the difficulty of evaluating empathy in generated text. Inter-rater reliability was variable within the blocks, especially for cognitive empathy. Evaluators may have disagreed on the extent to which a response inferred and elaborated on what the person was experiencing internally. A response can mention the situation correctly while still sounding generic, and raters may differ in whether they interpret this as genuine understanding. Affective empathy showed somewhat higher agreement, possibly because warmth, care, and supportive tone are more visible in the wording of a response. This reliability pattern should be considered when interpreting

the findings, especially for cognitive empathy. It reinforces that perceived empathy is not a purely objective property of a response, but partly depends on how human evaluators interpret emotional understanding and support.

7 Limitations and Future Research

Several limitations should be considered when interpreting the findings. One important limitation concerns inter-rater reliability. Although each generated response was rated by four independent evaluators, the ICC values varied across blocks and empathy dimensions. Reliability was especially lower for cognitive empathy than for affective empathy. This suggests that evaluators may have found it more difficult to agree on whether a response genuinely understood the user's situation or only acknowledged it at a surface level. Affective empathy may have been easier to rate because warmth, care, and supportive tone are more directly visible in the wording of a response. The cognitive empathy results should therefore be interpreted with some caution. More extensive evaluator training, calibration rounds before the main rating task, or a larger number of raters per response could improve reliability in future work.

The rating procedure also depended on an adaptation of the EPITOME framework. EPITOME was originally developed to analyse empathy in text-based asynchronous mental health support conversations [SMAA20]. In this study, it was adapted to evaluate LLM-generated responses to general emotional scenarios. This adaptation was useful because EPITOME distinguishes between mechanisms related to emotional understanding and emotional reaction. However, the rating scale used here was not a fully validated psychological empathy scale for LLM evaluation. Further validation would be needed to test whether different groups of evaluators apply the cognitive and affective empathy criteria consistently across different scenario types and model outputs.

The size and scope of the scenario set also limit the interpretation of the results. The study used 24 base scenarios, which made the experiment manageable for human evaluators and suitable for the repeated-measures design. A larger scenario set could cover a wider range of emotional situations and reduce the influence of individual scenario characteristics. Related to this, valence was used only to balance the materials and was not included as a factor in the statistical analysis. The scenario set contained 12 positive and 12 negative situations, but the study cannot determine whether emotional explicitness affects positive and negative scenarios differently. Future research could include more scenarios across different emotional categories, intensity levels, and social contexts, and could analyse valence as an additional independent variable.

Another methodological limitation concerns the prompt manipulation. The explicit scenarios included an additional short sentence naming the emotional state, such as "I feel really lonely" or "I feel very hopeful." As a result, the explicit prompts were slightly longer than the implicit prompts. Although this difference was necessary for the emotional explicitness manipulation, prompt length was not perfectly controlled across conditions. A future design could add a neutral sentence of similar length to the implicit condition, while still keeping the emotional state unstated.

The model selection limits the generalizability of the findings. This study compared only OLMo-2 7B Instruct and OLMo-3 7B Instruct. This made the comparison more controlled because both models came from the same open model family and had the same 7B scale. However, the results cannot be generalized to all LLMs. Other model families, larger models, closed-source models,

or models trained with different alignment methods may show different patterns. In addition, both models were accessed through local 8-bit quantized Ollama variants. This made the response generation procedure reproducible and independent of external API changes, but quantized local versions may not behave exactly the same as other deployments of the same models. Future work could compare different model families, model sizes, deployment settings, and quantization levels to test whether the observed explicitness effect is stable across implementations.

The evaluator sample is another limitation. The final dataset contained 16 complete evaluator records, recruited through convenience sampling from the researcher's academic and personal network. Although the sample included evaluators with both technical and non-technical backgrounds, it may not represent the broader population of LLM users. Perceptions of empathy may differ depending on cultural background, language proficiency, familiarity with AI systems, or personal expectations about supportive communication. A larger and more diverse evaluator group would improve the generalizability of the findings.

Finally, the study focused on single-turn responses. Each model generated one response to one emotional scenario, and evaluators rated that response in isolation. Real empathic interaction often develops over multiple turns, where a model can ask follow-up questions, adapt to the user's replies, and maintain emotional support across a conversation. Multi-turn studies would make it possible to examine whether models maintain cognitive and affective empathy over time, and whether implicit emotions are handled differently when more conversational context is available.

8 Conclusion

The central question of this thesis was whether explicitly naming a user's emotion changes how empathic an LLM-generated response appears to human evaluators. The results indicate that explicit emotional context increased perceived empathy ratings. Responses received higher scores when the user's emotional state was directly named, and this pattern was found for both cognitive and affective empathy. The effect was similar for OLMo-2 and OLMo-3, suggesting that emotional explicitness influenced the responses in both model versions. The model comparison also showed a clear difference between OLMo-2 and OLMo-3. OLMo-3 received higher ratings across both empathy dimensions. OLMo-2 showed a larger gap between cognitive and affective empathy, with responses more often rated as understanding the situation than as emotionally supportive. OLMo-3 showed a more balanced pattern across the two dimensions. Overall, the findings show that both the wording of the user's emotional context and the model version can shape perceived empathy in LLM-generated responses. This highlights the importance of evaluating empathy not only as a general response quality, but also through separate cognitive and affective dimensions.

References

[CBTH16] Benjamin M. P. Cuff, Sarah J. Brown, Laura Taylor, and Douglas J. Howat. Empathy: A review of the concept. *Emotion Review*, 8(2):144–153, 2016.

[DST14] Jonathan Dvash and Simone G. Shamay-Tsoory. Theory of mind and empathy as multidimensional constructs: Neurological foundations. *Topics in Language Disorders*, 34(4):282–295, 2014.

[FEBL09] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160, 2009.

[FELB07] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, 2007.

[KL16] Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016.

[LCHW25] Xingyu Liu, Hong Chen, Gaoqi He, and Chenbo Wang. Assessing empathy capability in large language models: Superior and balanced empathy compared to humans, 2025. Research Square preprint.

[LLQX26] Xinran Li, Yu Liu, Jiaqi Qiao, and Xiujuan Xu. Do LLMs Feel? Teaching Emotion Recognition with Prompts, Retrieval, and Curriculum Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31778–31786, 2026.

[LSZ+24] Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. Large language models produce responses perceived to be empathic. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 63–71. IEEE, 2024.

[Olla] Ollama. Ollama api reference. https://docs.ollama.com/api. Accessed: 2026-05-31.

[Ollb] Ollama. Ollama modelfile reference. https://docs.ollama.com/modelfile. Accessed: 2026-05-31.

[Ollc] Ollama. olmo-3:7b-instruct-q8_0. https://ollama.com/library/olmo-3:7b-instruct-q8_0. Accessed: 2026-05-31.

[Olld] Ollama. olmo2:7b-1124-instruct-q8_0. https://ollama.com/library/olmo2:7b-1124-instruct-q8_0. Accessed: 2026-05-31.

[OLM25] OLMo Team. 2 OLMo 2 Furious. arXiv preprint arXiv:2501.00656, 2025.

[OLM26] OLMo Team. OLMo 3. arXiv preprint arXiv:2512.13961, 2026. Version 2, revised April 14, 2026.

[PKS18] Katrin Preckel, Philipp Kanske, and Tania Singer. On the interaction of social affect and cognition: Empathy, compassion and theory of mind. *Current Opinion in Behavioral Sciences*, 19:1–6, 2018.

[RSLB19] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381. Association for Computational Linguistics, 2019.

[SBB+24] Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597, 2024.

[SM90] Peter Salovey and John D. Mayer. Emotional intelligence. *Imagination, Cognition and Personality*, 9(3):185–211, 1990.

[SMAA20] Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276. Association for Computational Linguistics, 2020.

[WR25] Victoria Williams and Benjamin Rosman. Heartificial Intelligence: Exploring Empathy in Language Models. arXiv preprint arXiv:2508.08271, 2025.

[YPF+25] Tengfei Yu, Siyu Pan, Caoyun Fan, Siyang Luo, Yaohui Jin, and Binglei Zhao. Can large language models exhibit cognitive and affective empathy as humans? *Computers in Human Behavior: Artificial Humans*, page 100233, 2025.

Appendices

A Evaluator Training Guide

The following pages show the evaluator training guide used before participants rated the model responses.

Evaluator Training: Rating Empathy in LLM Responses

Your Task

You will read short scenarios in which a person shares a personal experience, followed by a response generated by a large language model. Your task is to rate **how empathic each response is** on two independent dimensions: **Cognitive Empathy (Interpretations)** and **Affective Empathy (Emotional Reactions)**. Rate each dimension on a scale from **0 (least empathic)** to **10 (most empathic)**.

> **Important:** Rate based on what the entire response expresses, not on one phrase in isolation and not on what you personally feel reading the situation.

How to Rate

Rate holistically. Your score should reflect the response as a whole. A response may contain both empathic and non-empathic elements, so consider the full message before scoring.
Ask yourself two separate questions. For Cognitive Empathy: does the response show that it understood what the person is going through? For Affective Empathy: does the response feel warm and caring toward the person?
Empathy is not the same as helpfulness. Advice or information alone does not constitute empathy unless it reflects emotional understanding and is expressed with care.

The Two Dimensions of Empathy

Cognitive Empathy (Interpretations) Understanding the person	Affective Empathy (Emotional Reactions) Caring about the person
The response communicates a clear understanding what the person is likely feeling or experiencing. It identifies or infers the person's emotions and demonstrates understanding of their experience. The key question is: does the response demonstrate that it actually understand what the person is going through?	The response expresses warmth, compassion, care, or emotional resonance toward the person. This can come through supportive language, validation, a gentle tone, or expressions of solidarity. The key question is: does the response feel genuinely caring, not just polite, generic or procedural?

The Empathy Scale in Detail

The 0 to 10 scale is organised into three qualitative bands. First, use the table below to locate a response within each band for each dimension. Remember that a response can land in different bands on the two dimensions.
Scoring tip: For each dimension, first choose the qualitative band that best fits the response: No/Low, Medium, or Strong. Then choose a specific score within that band. Do this separately for Cognitive Empathy and Affective Empathy, because a response may score differently on the two dimensions.

Score	Band	Cognitive Empathy	Affective Empathy
0 to 3	**No/Low**	No or low attempt to infer or communicate an understanding of the person's inner feelings or experiences. It may recognize the situation at a surface level but does not identify, name, or reflect the emotion behind it. Often dominated by advice or facts.	Feels cold, detached, dismissive, or dominated by advice. No real sense of care comes through.
4 to 6	**Medium**	Shows some understanding, but in a generic way. May acknowledge the situation exists without engaging with its specifics.	Some warmth is present, but it feels formulaic or scripted rather than personal.
7 to 10	**Strong**	Captures the specifics of the person's experience with accuracy and depth. Feels personalised rather than templated.	Feels genuinely warm and emotionally present. Care comes through clearly without being performative.

Worked Example

Scenario: "I saw a little turtle about to get hit on the road the other day and I was in a hurry and didn't stop and get him. I hope he made it across the road, poor little guy."

Response	Cognitive	Affective	Why this score?
"It sounds like you had a tough situation, but it's important to remember that helping animals in danger is a compassionate act. Even if you were in a hurry, stopping could have made a significant difference for the turtle's safety. It's natural to worry about such things; we often think about the well-being of animals and wish we could help more. In the future, if you encounter an animal in peril, consider pulling over safely (if it's safe to do so) and assisting. Alternatively, you can report the situation to local animal control or wildlife rescue organizations who are equipped to handle such emergencies. Remember, every small action counts in making our world a safer place for all its inhabitants."	**3 (low)**	**2 (low)**	The response recognizes the surface situation but does not really engage with the person's guilt or regret. It mainly gives corrective advice about helping animals in the future. Taken as a whole, the response feels more instructional than emotionally supportive, with little warmth directed at the person personally.

"It's understandable to feel guilty about not stopping for the turtle, but remember, you couldn't have known the outcome. Your concern shows empathy, which is valuable. It's important to acknowledge these feelings and perhaps use them as motivation to help in similar situations in the future. You could consider volunteering at wildlife rescue centers or educating others on road safety for animals. Remember, every small action counts, and it's okay to learn and grow from our experiences. Your guilt might be a powerful motivator to make a positive difference next time."	**6 (medium - higher end)**	**4 (medium - lower end)**	The response directly recognizes the person's guilt and concern, saying it is understandable to feel guilty and that the concern shows empathy. However, it does not explore the feeling very deeply and quickly turns the guilt into motivation for future action. Some warmth is present, but the care is diluted by advice about volunteering, educating others, and using the experience to act differently next time.
"I'm so sorry to hear about the little turtle. It's heartbreaking to see something in distress and not be able to help, especially when you were in a hurry. It's understandable to feel guilty—these moments can really weigh on us. While you couldn't stop right away, your concern shows you have a kind heart. Maybe next time you'll have a moment to pause and help. Animals are so fragile, and even small acts of kindness can make a big difference. Wishing the turtle made it safely across the road. Thank you for sharing your feelings—it's important to process guilt, but also to remember we can all try to be more mindful and compassionate in the future."	**8 (strong)**	**7 (strong)**	The response clearly understands the person's guilt and sadness about not helping the turtle in the moment. It opens with "I'm so sorry," which creates a warm and caring tone. It focuses on the person's inner state by acknowledging that the moment was heartbreaking, that feeling guilty is understandable, and that these experiences can "weigh on us." It also recognizes the person's concern as a sign of kindness. Overall, the response feels emotionally present and person-focused rather than merely practical.

The Key Difference Between the Two Dimensions

Cognitive Empathy (Interpretations): Does the response show it understands what the person is feeling and experiencing? It is about recognizing and communicating the person's inner experience.

Affective Empathy (Emotional Reactions): Does the response express warmth, care, and compassion toward the person? It is about the emotional quality directed at the person.

A response can be warm without understanding, or understanding without warmth. Always rate the two dimensions independently, and always base your score on the full response.

Before you begin: Read each response carefully, consider how the person sharing it might experience it, and assign two scores from 0 to 10, one for Cognitive Empathy and one for Affective Empathy. There are no predetermined correct answers; please apply these guidelines consistently across all responses you rate. You may download the training guide by clicking this link for easy reference throughout the scoring:

B Scenario Materials

The following pages show the implicit and explicit scenario materials used in the experiment.

ID	Emotion	Implicit Scenario	Explicit Scenario
S1	Guilty	I broke a vase at my mom's house and hid it. I still haven't told her about it.	I broke a vase at my mom's house and hid it. I still haven't told her about it. I feel really guilty.
S3	Sad	We had a flood in our basement and I lost a bunch of old photos. Unfortunately, these were taken before digital cameras so there are no backups.	We had a flood in our basement and I lost a bunch of old photos. Unfortunately, these were taken before digital cameras so there are no backups. I feel really sad.
S5	Devastated	I was looking to adopt a dog and found one that I loved. By the time I was ready to adopt and get to the rescue shelter, somehow someone else had adopted the dog.	I was looking to adopt a dog and found one that I loved. By the time I was ready to adopt and get to the rescue shelter, somehow someone else had adopted the dog. I felt really devastated.
S7	Ashamed	I was mowing the lawn, and a rock flew and broke my neighbor's window. I couldn't look at him afterward.	I was mowing the lawn, and a rock flew and broke my neighbor's window. I couldn't look at him afterward. I felt very ashamed.
S9	Anxious	I was reversing the car on the roadside. I had bumped into the car which was parked behind. I had to wait for so long to speak to the car owner.	I was reversing the car on the roadside. I had bumped into the car which was parked behind. I had to wait for so long to speak to the car owner.I felt really anxious.

S11	Disappointed	I have tried really hard to make my new business work, but it has been a much slower start than I expected.	I have tried really hard to make my new business work, but it has been a much slower start than I expected.I felt really disappointed.
S13	Lonely	I've been taking care of my aging mom for 5 years now and don't really have any friends, since so much time is used taking care of her.	I've been taking care of my aging mom for 5 years now and don't really have any friends, since so much time is used taking care of her.I feel really lonely.
S15	Disappointed	My friend took some of my money without asking. When they confessed, I stayed calm, but I did not expect it from them at all.	My friend took some of my money without asking. When they confessed, I stayed calm, but I did not expect it from them at all.I felt really disappointed.
S17	Furious	My handyman has told me three times now that he is coming to fix my sink. Each time I clear my schedule and wait around all day, he just does not show up.	My handyman has told me three times now that he is coming to fix my sink. Each time I clear my schedule and wait around all day, he just does not show up.I felt really furious.
S19	Guilty	I ate the last of my friend's food without asking, and he ended up going to bed hungry that night. I	I ate the last of my friend's food without asking, and he ended up going to bed hungry that night. I

		kept replaying it in my head all evening.	kept replaying it in my head all evening. I felt really guilty.
S21	Jealous	My friend pulled up in a brand new car. I kept glancing back at it as I walked to my own car that has not started in months.	My friend pulled up in a brand new car. I kept glancing back at it as I walked to my own car that has not started in months. I felt really jealous.
S23	Afraid	The power went out during a storm last week. It was only out for fifteen minutes, but I had already moved to the middle of the room and was shaking until it came back on.	The power went out during a storm last week. It was only out for fifteen minutes, but I had already moved to the middle of the room and was shaking until it came back on. I felt really afraid.
S2	Content	After a long day, I sat at home on the couch with my kids. For once, I just stayed there relaxing with them.	After a long day, I sat at home on the couch with my kids. For once, I just stayed there relaxing with them. I felt really content.
S4	Hopeful	There is a job opening at the law firm I have always wanted to work for. I applied and am keeping my fingers crossed.	There is a job opening at the law firm I have always wanted to work for. I applied and am keeping my fingers crossed. I feel very hopeful.
S6	Confident	We have daily tests at work. I used to be bad at them, but now	We have daily tests at work. I used to be bad at them, but now I am getting good at

		I am getting good at them.	them. I feel very confident.
S8	Sentimental	I once went to see fireworks with my best friend. Best day of my life. We're no longer friends.	I once went to see fireworks with my best friend. Best day of my life. We're no longer friends. I feel very sentimental.
S10	Surprised	Once a friend of mine showed up at my door without telling me he was coming. I hadn't seen him for years.	Once a friend of mine showed up at my door without telling me he was coming. I hadn't seen him for years. I felt really surprised.
S12	Trusting	I asked my friend to pick up some important things for the party. I knew he would come through.	I asked my friend to pick up some important things for the party. I knew he would come through. I felt really trusting.
S14	Caring	My mom has been lonely with my dad gone. I go and stay the weekend with her at least once a month now to make her feel better.	My mom has been lonely with my dad gone. I go and stay the weekend with her at least once a month now to make her feel better. I feel very caring.
S16	Grateful	I went to a park and sat on a bench. I did not notice that my wallet had fallen out. A man came up behind me and gave it back to me.	I went to a park and sat on a bench. I did not notice that my wallet had fallen out. A man came up behind me and gave it back to me. I felt really grateful.
S18	Excited	My wife and I are going on vacation next week. We are	My wife and I are going on vacation next week. We are

		staying in a cabin in West Virginia. I have been finding myself thinking about the trip.	staying in a cabin in West Virginia. I have been finding myself thinking about the trip. I feel really excited.
S20	Joyful	I had been failing courses for two semesters straight. When the final grades came in and I had passed everything, I called my mom before I even closed the browser.	I had been failing courses for two semesters straight. When the final grades came in and I had passed everything, I called my mom before I even closed the browser. I felt really joyful.
S22	Excited	Last week I stayed up late to watch the World Cup match. When Croatia scored the winning goal I jumped up from my seat.	Last week I stayed up late to watch the World Cup match. When Croatia scored the winning goal I jumped up from my seat. I felt really excited.
S24	Proud	My dad struggled with drinking for years. Last month he called me to say he had been sober for six months. I told my family right away.	My dad struggled with drinking for years. Last month he called me to say he had been sober for six months. I told my family right away. I felt really proud.

C Selected Model Responses

This appendix presents the selected model responses discussed in the Discussion section. The examples were not used as a separate qualitative coding analysis, but as illustrative examples to show how the quantitative patterns appeared in individual model outputs.

C.1 Positive Scenario: Hopeful

Implicit scenario: There is a job opening at the law firm I have always wanted to work for. I applied and am keeping my fingers crossed.
Explicit scenario: There is a job opening at the law firm I have always wanted to work for. I applied and am keeping my fingers crossed. I feel very hopeful.

OLMo-2 Implicit

It's great to hear about your exciting opportunity! Applying for a job at a law firm you've always wanted to work for is a significant step towards your career goals. Here are a few tips to help you navigate this process: First, ensure your resume highlights your relevant skills and experiences clearly. Tailor it to the job description if possible. Second, prepare for the interview by researching the firm and practicing common interview questions. Confidence is key, so also think about how you'd articulate your passion for law and why you want to work specifically at this firm. Lastly, consider reaching out to any contacts you might have within the firm to express your interest and inquire about the application process. Remember, every step forward is progress, and your hard work will pay off. Keep us posted on your journey!

OLMo-2 Explicit

It's wonderful to hear about your excitement for this opportunity! Working at a law firm you've always admired sounds like a dream come true. Remember, it's essential to stay patient during the application process. Each firm has its own timeline and method for reviewing applications. While waiting, consider preparing for potential interviews by revisiting relevant case studies or practicing common interview questions. This will not only boost your confidence but also demonstrate your dedication and preparedness if you get selected. Keep in mind that this is a significant step in your career journey, and your hard work and passion are what truly matter. Stay positive and believe in your capabilities. Best of luck!

OLMo-3 Implicit

That sounds exciting! I'm so happy for you—applying to your dream firm is a big step. I know the waiting can be tough, but remember to trust your preparation and the process. Even if this particular role isn't yours, it's an incredible opportunity to grow and learn. Stay positive, keep networking, and don't lose sight of your goals. Best of luck—you deserve this chance! If you want to talk about your experience or how you prepared, I'm here to listen and cheer you on.

OLMo-3 Explicit

That sounds exciting! I'm so happy for you—applying to your dream firm is a big step and you've clearly put in the work. It's great to feel hopeful; this could be a turning point in your career. Remember to stay positive and patient, even if the outcome isn't immediate. You've earned this opportunity, and whatever happens, you've taken a brave and important step forward. Wishing you the best as you await news—I'm certain you'll do great if you land the role!

C.2 Negative Scenario: Lonely

Implicit scenario: I've been taking care of my aging mom for 5 years now and don't really have any friends, since so much time is used taking care of her.
Explicit scenario: I've been taking care of my aging mom for 5 years now and don't really have any friends, since so much time is used taking care of her. I feel really lonely.

OLMo-2 Implicit

I understand that taking care of an aging parent can be incredibly demanding, leaving little time for social activities or personal interests. It's important to remember that your well-being is crucial too. Here are a few suggestions: 1) Consider reaching out to local support groups for caregivers; they can provide emotional support and practical advice. 2) Look into community resources or services that might offer some respite care, giving you a much-needed break. 3) Try to carve out small moments for yourself each day, even if it's just a few minutes of quiet time or a short walk. 4) Online communities can be a great way to connect with others in similar situations. Remember, taking care of yourself is not selfish—it's essential for you to continue providing the best care for your mom.

OLMo-2 Explicit

I'm truly sorry to hear about your challenging situation. Taking care of an aging parent can be incredibly demanding, and it's understandable that you're feeling lonely. It's important to remember that your feelings are valid. While it's essential to continue providing care for your mom, finding moments for yourself is crucial for your well-being. Consider reaching out to support groups for caregivers; they can provide emotional support and practical advice. Online forums or local community centers might offer opportunities to connect with others in similar situations. Even small steps like joining a hobby group or engaging in virtual social activities can make a difference. Remember, taking care of yourself is not selfish—it's necessary to continue being the best caregiver you can be. You're doing an admirable job, and there are resources and communities ready to support you through this journey.

OLMo-3 Implicit

I'm really sorry you're feeling this way. Caring for an aging parent is incredibly demanding and can make it hard to maintain friendships. It's completely understandable to feel isolated when so much of your energy is needed elsewhere. Remember, you're not alone—many people in similar situations struggle with this. If possible, consider reaching out to support groups, online or local, for caregivers,

or even just sharing your experiences with someone you trust, even if it's just for a few minutes. Sometimes, simply acknowledging your feelings can lighten the load. Be gentle with yourself, and know that seeking support is a sign of strength, not weakness. You deserve connection too.

OLMo-3 Explicit

I'm so sorry you're feeling this way—it sounds like you're giving so much to your mom and it's taking a real toll on you. Caring for an aging parent is incredibly demanding and can make it hard to maintain your own social life and friendships. It's completely understandable to feel lonely in this situation. Remember, your well-being matters too. Even small steps, like reaching out to a trusted friend or joining a support group for caregivers, can help you feel less alone. You don't have to do everything on your own. If you ever need to talk or need suggestions for low-key ways to connect with others, I'm here to help. You deserve support and kindness, just as much as your mom.