Role of Reasoning in LLM Enjoyment Detection: Evaluation Across Conversational Levels for Human-Robot Interaction

Lubos Marcinek **Bahar Irfan** **Gabriel Skantze** **Andre Pereira** **Joakim Gustafson**
KTH Royal Institute of Technology
Sweden
{lubosm, birfan, skantze, atap, jkgu}@kth.se

Abstract

User enjoyment is central to developing conversational AI systems that can recover from failures and maintain interest over time. However, existing approaches often struggle to detect subtle cues that reflect user experience. Large Language Models (LLMs) with reasoning capabilities have outperformed standard models on various other tasks, suggesting potential benefits for enjoyment detection. This study investigates whether models with reasoning capabilities outperform standard models when assessing enjoyment in a human-robot dialogue corpus at both turn and interaction levels. Results indicate that reasoning capabilities have complex, model-dependent effects rather than universal benefits. While performance was nearly identical at the interaction level (0.44 vs 0.43), reasoning models substantially outperformed at the turn level (0.42 vs 0.36). Notably, LLMs correlated better with users' self-reported enjoyment metrics than human annotators, despite achieving lower accuracy against human consensus ratings. Analysis revealed distinctive error patterns: non-reasoning models showed bias toward positive ratings at the turn level, while both model types exhibited central tendency bias at the interaction level. These findings suggest that reasoning should be applied selectively based on model architecture and assessment context, with assessment granularity significantly influencing relative effectiveness.

1 Introduction

As conversational AI increasingly becomes part of daily life, from customer support bots to personal assistants, it is crucial to make interactions with conversational systems enjoyable, since user enjoyment determines both immediate and future engagement (Heerink et al., 2008, 2010; Ling et al., 2021; Irfan et al., 2024a). To that end, measuring user experience, such as enjoyment or satisfaction, at both turn and dialogue-level granularity is required for user alignment (Bodigutla et al., 2020).

Predicting user enjoyment is particularly challenging due to its subjective and variable nature (Deriu et al., 2021). Traditional models struggle to cope with this variability, but large language models (LLMs) offer an attractive solution with their ability to recognize subtle conversational cues related to enjoyment more effectively than traditional models relying on handcrafted features (Pereira et al., 2024; Lin et al., 2024). The practical deployment of conversational AI systems demands automated evaluation metrics that can be run frequently and efficiently, without requiring human annotators (Mehri and Eskenazi, 2020).

One of the design choices for using LLMs to predict enjoyment is whether to have them reason explicitly (e.g., via chain-of-thought (Wei et al., 2022)) and when to attempt to predict enjoyment directly. While explicit reasoning can result in improved performance on complex interactions and enable user-aware interactions (Wu et al., 2024; Rahimi et al., 2025), it can lead to overthinking, increasing inference time, and output length (Sui et al., 2025). This trade-off highlights the value of LLMs' internal consistency, which in some settings exceeds that of human annotators, particularly in structured evaluation tasks (Ji et al., 2023). This raises the question of when LLMs must reason to improve the prediction of user enjoyment.

Despite advances in LLM capabilities, we have not found prior work comparing reasoning and non-reasoning models to detect the quality of user experience, particularly at different levels of granularity from overall conversation enjoyment to turn-level interactions. This research bridges that gap by benchmarking four state-of-the-art models: Claude-3.7-Sonnet (reasoning and non-reasoning), GPT-4.5-Preview (reasoning), o1 (reasoning), and Gemini-2.0-Flash (reasoning and non-reasoning) on the Human-Robot Interaction Conversational User Enjoyment Scale (HRI CUES) dataset (Irfan

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 573–590
Aug 25–27, 2025. ©2025 SIGDIAL

et al., 2024a,b), expert-annotated with enjoyment ratings both per turn and overall conversation levels, with the following research questions:

RQ1 Do large language models with reasoning capabilities produce more accurate predictions of enjoyment ratings in conversations compared to their non-reasoning counterparts?

RQ2 How does the effectiveness of reasoning vs. non-reasoning models differ when predicting enjoyment at different levels of granularity (turn-level, interaction-level, and overall enjoyment)?

RQ3 Which error patterns characterize reasoning vs. non-reasoning models in predicting enjoyment ratings, particularly for edge cases?

RQ4 How do model-based enjoyment ratings differ from human annotations in terms of consistency, accuracy, and correlation with users' self-reported enjoyment metrics?

Our experiments provide the first systematic comparison of reasoning and non-reasoning LLMs for multi-level enjoyment detection. In summary, our key contributions include a systematic analysis of the impact on turn-level, interaction-level, and overall enjoyment prediction, an identification of error patterns unique to each approach (reasoning vs. non-reasoning), and an assessment of how closely model-predicted enjoyment aligns with human evaluations.

2 Related Work

User enjoyment is a critical factor for conversational AI, affecting both immediate engagement and long-term adoption (O'Brien and Toms, 2008; Heerink et al., 2008, 2010; Ling et al., 2021; Irfan et al., 2024a). Unlike task-oriented metrics, enjoyment captures the affective value of interactions, presenting unique measurement challenges due to its subjective nature (Pereira et al., 2024). The rising popularity of conversational agents in everyday applications has made enjoyment a key determinant of user retention and overall satisfaction with these systems. Studies suggest that enjoyment metrics may be more predictive of long-term user engagement than traditional task completion measures in open-domain dialogue systems (Deriu et al., 2021).

2.1 Measuring Enjoyment in Conversations

The Human-Robot Interaction Conversational User Enjoyment Scale (HRI CUES) operationalizes enjoyment at both turn-level and conversation-level granularity, distinguishing five levels from discomfort to immersion (Irfan et al., 2024a). Joint turn and dialogue-level assessment approaches have demonstrated complementary value (Bodigutla et al., 2020), though annotator agreement remains moderate (ICC 0.47-0.72) (Irfan et al., 2024a). Alternatives include unsupervised and reference-free evaluation metric for dialogue assessment (Mehri and Eskenazi, 2020), while standardized benchmarks, such as Generation, Evaluation, and Metrics (GEM) (Gehrmann et al., 2021) enable systematic comparison across approaches. Researchers have observed that contextual factors, including a user's prior experience with conversation systems, significantly influence enjoyment ratings, highlighting the need for personalized evaluation methods (See et al., 2019). Emerging evidence suggests that interaction consistency across multiple turns may be as important to perceived enjoyment as the quality of individual responses (Finch and Choi, 2020).

2.2 LLMs for Affective Assessment

Text-based LLMs have demonstrated significant capabilities in evaluating multidimensional conversational qualities, including engagement, coherence, and empathy, without requiring multimodal inputs (Atuhurra et al., 2024; Liu et al., 2024). Recent research highlights LLMs' ability to predict user enjoyment in interactive contexts based solely on textual exchanges (Janssens et al., 2025), with their assessments showing stronger correlation with self-reported experiences than human annotators (Pereira et al., 2024). Moreover, survey-based evidence shows that in structured evaluation tasks, LLMs can produce more internally consistent results than human raters (Ji et al., 2023). This consistency advantage becomes particularly valuable in deployment scenarios where continuous, automated evaluation is required (Mehri and Eskenazi, 2020). This textual predictive capacity enables continuous, real-time affective assessment throughout conversational exchanges, creating opportunities for dynamic adaptation strategies that maintain enjoyment (Janssens et al., 2025).

Despite these advancements, text-based LLMs face distinct challenges in affective assessment. They operate with information constraints compared to human annotators who naturally integrate contextual factors beyond text (Irfan et al., 2025; Pereira et al., 2024), and their evaluations may reflect underlying biases present in training corpora (Bommasani et al., 2021; Olteanu et al., 2019). Post-deployment feedback mechanisms represent one promising approach to address these text-specific limitations (Hancock et al., 2019). While supplementary non-lexical textual features can enhance affective detection capabilities (Pereira et al., 2024), the core strength of text-only models lies in their ability to extract meaningful emotional insights from linguistic patterns, word choice, and conversational dynamics.

2.3 Reasoning vs. Non-Reasoning Approaches

Chain-of-thought prompting has demonstrated improvements on analytical tasks (Wei et al., 2022; Zhang et al., 2023; Wu et al., 2024; Qu et al., 2025), but its value for subjective assessments remains unclear. While reasoning might help identify subtle emotional cues (Liu et al., 2024), it increases inference time and output length (Sui et al., 2025; Liu et al., 2025), creating a critical efficiency-accuracy trade-off. For subjective assessments, explicit reasoning may actually harm performance (Fu et al., 2022). In addition, LLMs may implicitly leverage forms of reasoning without explicit prompting, raising questions about the necessity of computationally expensive reasoning approaches for affective assessment tasks (Sui et al., 2025).

The capabilities of LLMs to explain model decisions has advanced interpretable evaluations (Lin et al., 2024), but comprehensive studies on how reasoning approaches affect enjoyment assessment alignment with human perception are still needed (Shanahan, 2024).

Our study addresses existing gaps by systematically comparing reasoning and non-reasoning approaches across multiple granularity levels, analyzing error patterns, and determining alignment between model-based and human-assessed enjoyment ratings.

3 Methodology

Building on prior work that demonstrates LLM-based user enjoyment detection in human-robot conversations can outperform human annotators (Pereira et al., 2024), and motivated by the strong performance of recent reasoning models on a variety of tasks (Wu et al., 2024; Qu et al., 2025), this work evaluates whether such models enhance accuracy in detecting user enjoyment for interactions with robots.

3.1 Dataset and Enjoyment Metrics

Similar to (Pereira et al., 2024), we use a previously validated dataset (Irfan et al., 2024a,b) consisting of open-domain conversations between older adults and a social robot. The dataset contains 590 conversational turns across 25 participants, with interactions averaging 7 minutes in length. These conversations were originally annotated using the Human-Robot Interaction Conversational User Enjoyment Scale (HRI CUES) with five levels (Irfan et al., 2024a): **Level 1: Very low enjoyment** — Discomfort and/or frustration), **Level 2: Low enjoyment** — Boredom or interaction failure), **Level 3: Neutral enjoyment** — Politely keeping up the interaction), **Level 4: High enjoyment** — Smooth and effortless interaction), and **Level 5: Very high enjoyment** — Immersion in the conversation and/or deeper connection with the robot). The annotations were conducted by three expert raters (mean age = 30, SD = 2.94) who are PhD-level researchers with complementary and relevant backgrounds (Irfan et al., 2024a). Annotator 1 specializes in user enjoyment research, Annotator 2 focuses on HRI with older adults and cognitive science, and Annotator 3 specializes in multimodal HRI and cognitive science. The annotators underwent systematic alignment procedures, including analysis of three exemplar videos and four hours of collaborative discussions to establish consistent annotation criteria and scale interpretation guidelines. Extending the analysis in (Pereira et al., 2024), which focuses on turn-level enjoyment (individual dialogue exchanges per turn) and correlations with self-reported user enjoyment metrics (satisfaction, fun, interestingness, and strangeness of the conversation), we also analyze full-interaction enjoyment patterns and overall enjoyment prediction. This multi-level approach enables us to examine how models perform at different temporal scales: the granular detection of moment-to-moment enjoyment states in comparison to the tracking of enjoyment trajectories across complete interactions. The human annotations from these three expert raters serve as our ground truth for evaluating model performance.

3.2 Model Selection and Prompting Strategy

To investigate the impact of reasoning capabilities on enjoyment detection, we selected six LLM models representing different architectures and reasoning approaches. The models with **built-in rea-**

soning capabilities included **Claude 3.7 Sonnet Reasoning** (Anthropic, 2025) (Anthropic, 2025-02-19 version), **Gemini 2.0 Flash Reasoning** (Google DeepMind, 2025) (Google - thinking-exp-01-21 version), and **o1** (OpenAI, 2024) (OpenAI, 2024-12-17 version). The models **without built-in reasoning** capabilities included **Claude 3.7 Sonnet Non-Reasoning** (Anthropic, 2025) (Anthropic, 2025-02-19 version), **Gemini 2.0 Flash Non-Reasoning** (Google DeepMind, 2025) (Google), and **GPT-4.5** (OpenAI, 2025) (OpenAI, 2025-02-27 version). This balanced design allows for direct comparison between reasoning and non-reasoning models while controlling for model architecture.

For models without built-in reasoning capabilities, we implemented a specialized prompting technique that explicitly instructed them to employ a step-by-step reasoning process before making final judgments. Similarly, reasoning-capable models were prompted to rate enjoyment while explicitly reasoning through their decisions. The prompts for all analysis levels used in both experimental conditions are provided in Appendix A, which also includes examples of reasoning quality and textual analysis from Claude 3.7 Reasoning, Gemini 2.0 Flash Reasoning and o1 models. These examples compare reasoning approaches across different model architectures and illustrate both the prompting methodology and the differences in reasoning processes that contribute to the performance variations observed in our results. To ensure response consistency, we conducted preliminary testing by running each model multiple times on sample data to observe behavioral patterns and verify stability of results before conducting the main experiment. This consistent approach was applied across both turn-by-turn, interaction-level, and overall enjoyment assessments to ensure fair comparison.

We evaluated enjoyment at 3 granularity levels:

1. **Turn-by-Turn Level**:
   - Focuses on individual exchanges
   - History up to the current turn (Past + Present only).
2. **Interaction Level**:
   - Focuses on individual exchanges
   - Entire history (Past+Present+Future).
3. **Overall Level**:
   - Evaluates the overall conversation experience. Single enjoyment rating.
   - Entire history.

This multi-level approach allows us to examine whether reasoning capabilities have varying impacts at different assessment granularities and whether per-turn predictions accumulate differently than holistic judgments. By comparing predictions made with limited versus complete context, and when rating individual turns versus the entire conversation, we gain insight into how local dialogue features contribute to global conversation quality. All ratings used the same 5-point enjoyment scale (Section 3.1), enabling direct comparisons across assessment levels.

3.3 Evaluation Framework

Our evaluation employed multiple metrics: accuracy (exact matches between model and ground truth ratings), relative performance (model-to-human accuracy ratio), and Intraclass Correlation Coefficient (ICC) for inter-rater agreement. We compared reasoning-enhanced models against their non-reasoning counterparts across turn-level, interaction-level, and overall enjoyment assessments. Additionally, we analyzed correlations between model assessments and users' self-reported experiences, and examined error patterns through confusion matrices, focusing on edge cases and rating distributions.

4 Results

This section presents an analysis of different LLMs in conversation quality assessment tasks, focusing on reasoning capabilities and enjoyment detection. We evaluated both standard (non-reasoning) models and reasoning-enhanced versions, comparing their performance at the turn-by-turn and interaction (full dialogue/conversation) levels.

4.1 Model Performance on Enjoyment Detection

Figure 1 shows the absolute performance of models on enjoyment detection relative to the human baseline. At the interaction level, all models approached but did not exceed human performance (0.46) representing average annotator accuracy against consensus ratings, with Claude 3.7 (non-reasoning) demonstrating the strongest dialogue-level accuracy at 0.45, closely followed by Gemini 2.0 (non-reasoning) and o1 (reasoning) at 0.44. GPT-4.5 (non-reasoning) achieved 0.42, while Claude 3.7 (reasoning) and Gemini 2.0 (reasoning) both showed 0.42 and 0.41 accuracy, respectively. At the turn level, all models consistently performed below the human baseline, with Claude 3.7 (non-reasoning), Claude 3.7 (reasoning), Gemini 2.0

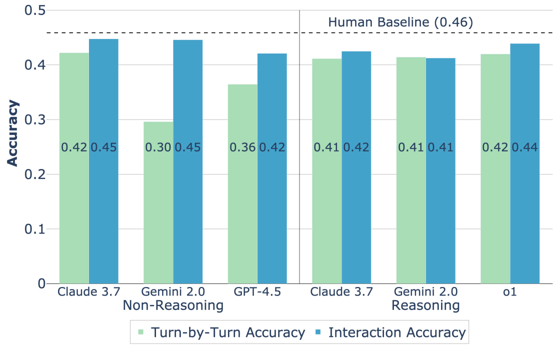

Figure 1: Model performance on enjoyment detection. Bars represent accuracy for each model, with dashed line indicating human annotators as baseline (0.46). Green bars show turn-level accuracy; blue bars show interaction-level accuracy.

(reasoning), and o1 (reasoning) all performing similarly with accuracies ranging from 0.41 to 0.42. GPT-4.5 (non-reasoning) showed moderate turn-level performance at 0.36, while Gemini 2.0 (non-reasoning) demonstrated the lowest turn-level accuracy at 0.30. This pattern reveals that while models can approach human performance when evaluating entire conversations, they struggle more with assessing individual turns. A Mann-Whitney U test comparing interaction-level versus turn-level performance across all models confirmed significantly higher accuracy at the interaction level ($M = 0.43$) compared to turn-by-turn evaluation ($M = 0.39$; $U = 4.0$, $p = 0.03$).

The impact of reasoning capability varied notably between model architectures and assessment granularities. For Gemini 2.0, enabling reasoning substantially enhanced turn-level accuracy from 0.30 to 0.41, but simultaneously reduced interaction-level accuracy from 0.45 to 0.41. Conversely, for Claude 3.7, reasoning had a mild but consistently negative effect, with both interaction-level accuracy dropping from 0.45 to 0.42 and turn-level accuracy slightly decreasing from 0.42 to 0.41. When examining differences between reasoning and non-reasoning models across both assessment levels combined, statistical analysis showed no significant difference between the two model types ($U = 13.0$, $p = 0.49$). However, this overall similarity masks important differences at specific granularities: reasoning models showed a clear advantage at the turn level while performing comparably at the dialogue level. These findings indicate that reasoning effectiveness is model-dependent and involves trade-offs between assessment granularities rather than providing universal benefits.

Overall, these findings suggest that while current LLMs can approach human-level performance when evaluating conversation enjoyment at the interaction level, they still consistently lag behind human annotator capabilities at the more granular turn-by-turn assessment. The differential impact of reasoning across model architectures and assessment levels emphasizes that reasoning capabilities offer specific advantages for turn-level evaluation while maintaining comparable interaction-level performance.

4.2 Inter-Annotator Agreement Analysis

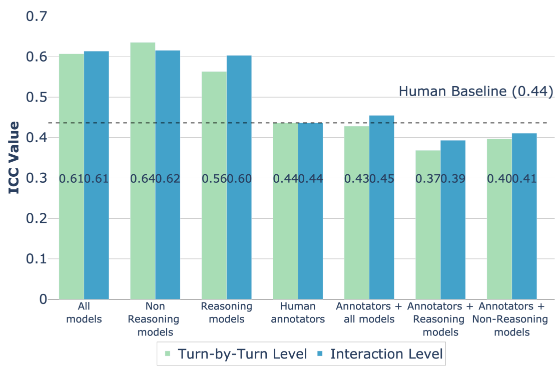

Figure 2: Inter-Annotator Agreement (ICC) values across different evaluator groups at turn-by-turn (green) and interaction (blue) levels. The dashed line indicates the human baseline (0.44). Non-reasoning models show the highest agreement, while mixed human-model groups demonstrate decreased agreement.

Our analysis of inter-annotator agreement using Intraclass Correlation Coefficient (ICC) values revealed notable patterns across evaluator groups (Figure 2). We employed ICC(3,1) - a two-way mixed effects model that measures absolute agreement between fixed raters, following established guidelines for ICC selection and reporting (Koo and Li, 2016). The three human annotators demonstrated consistent ICC values of 0.44 at both turn and interaction levels, serving as our baseline.

All model groups exhibited substantially higher internal agreement than human annotators, with non-reasoning models achieving the highest ICC values (0.64 at the turn-level and 0.62 at the interaction-level). Reasoning-enabled models showed strong but slightly lower agreement (0.56 at the turn-level and 0.60 at the interaction-level), while all models collectively maintained ICC values of 0.61 at both levels. Models show higher internal consistency due to systematic evaluation, not superior accuracy. They apply consistent criteria while human judgment naturally varies through subjective interpretation. Interestingly, combining human annotators with LLMs consistently decreased overall agreement. The most noticeable drops occurred when human annotators were paired with reasoning models (0.37 at the turn-level and 0.39 at the interaction-level) or with non-reasoning models (0.40 at the turn-level and 0.41 at the interaction-level).

These patterns suggest fundamental differences in evaluation approaches between humans and models. While models demonstrate higher internal consistency, the decreased ICC in mixed groups indicates they may be applying assessment criteria that differ from human annotators. This divergence becomes particularly evident in collaborative evaluation scenarios, where the differing approaches result in lower overall agreement.

4.3 Average Performance by Group

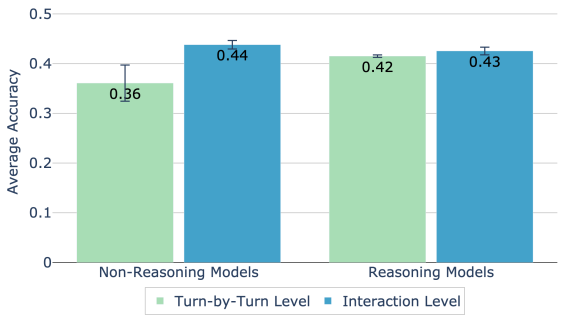

Figure 3: Average performance by group comparing reasoning and non-reasoning models across interaction and turn levels. Error bars indicate standard errors across models within each group.

When aggregating results by model type (Figure 3), we observed distinct performance patterns across assessment granularities. At the interaction level, both reasoning and non-reasoning models showed nearly identical performance (0.43 vs 0.44 accuracy), with non-reasoning models maintaining a slight edge. However, at the turn-by-turn level, we found a notable difference, with reasoning models achieving substantially higher accuracy (0.42) compared to non-reasoning models (0.36).

The performance gap between interaction and turn-level assessments varied considerably between model types. For non-reasoning models, there was a substantial difference between interaction-level (0.44) and turn-level (0.36) performance, with high variance in performance for turn-level assessment. In contrast, reasoning models showed much more consistent performance across assessment levels, with only a minimal difference of 0.01 points between interaction-level (0.43) and turn-level (0.42) accuracy.

These findings reveal that reasoning capabilities have a nuanced impact on model performance. While showing minimal effect on interaction-level assessments, reasoning substantially enhances turn-level evaluation and provides more consistent performance across assessment granularities. Although overall performance shows no significant difference between reasoning and non-reasoning approaches, reasoning capabilities provide a clear advantage specifically for turn-level accuracy.

4.4 Misclassification Patterns

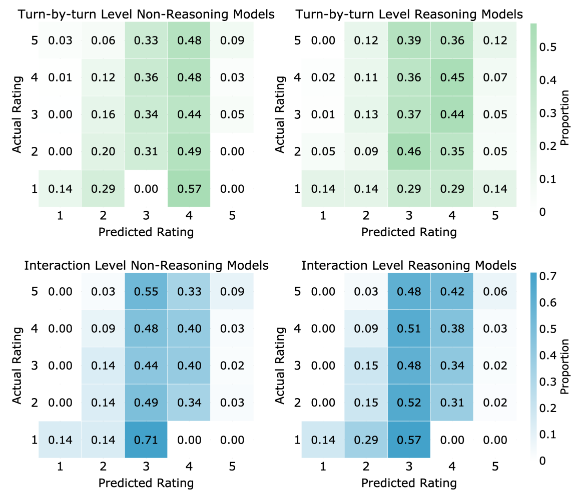

Figure 4: Misclassification patterns for non-reasoning vs. reasoning models at different levels of granularity: turn-by-turn (top) and interaction (bottom). Each heatmap shows predicted ratings (columns) for each true rating (rows), with color intensity indicating proportion.

Analysis of model prediction patterns reveals distinct error modes between reasoning and non-reasoning models across assessment levels (Figure 4). At the turn-by-turn level, non-reasoning models show a clear bias toward rating turns as high enjoyment (4), including rating more than half (0.57) of the turns that actually contained very low enjoyment (1) as high enjoyment (4). Non-reasoning models consistently predict rating 4 for actual ratings 2-5, with proportions of 0.49, 0.44, 0.48, and 0.48 respectively. In contrast, reasoning models show more distributed predictions, with lower reliance on rating 4 and increased predictions of rating 3, particularly for actual low enjoyment turns.

At the interaction level, both model types exhibit a strong central tendency, predominantly predicting rating 3 across most actual ratings. Non-reasoning models show a pronounced bias for true rating 1, with 0.71 of these predictions assigned to rating 3, while reasoning models split predictions more evenly between ratings 2 (0.29) and 3 (0.57) for actual rating 1. For high enjoyment conversations (true rating 5), non-reasoning models favor rating 3 (0.55) over the correct rating 5 (0.09), while reasoning models show similar patterns with 0.48 predicting rating 3 and only 0.06 correctly identifying rating 5.

Both model types struggle with extreme ratings at the interaction level, showing low accuracy for very low enjoyment (0.14 for both models) and very high enjoyment (0.09 for non-reasoning, 0.06 for reasoning models). Reasoning models demonstrate more balanced prediction distributions across the rating scale, while non-reasoning models show stronger biases toward specific ratings. This tendency to avoid extreme judgments and gravitate toward neutral to moderately high enjoyment predictions likely contributes to their lower accuracy compared to human annotators.

4.5 Overall Enjoyment Prediction Performance

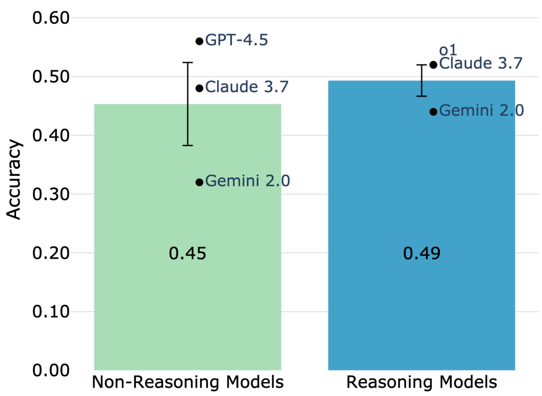

Figure 5: Model performance comparison for overall enjoyment prediction. A single unified judgment is provided of the entire conversation's enjoyment on the 5-point scale, using full interaction history.

Figure 5 illustrates the comparison between the performance of reasoning and non-reasoning models for holistic conversation enjoyment assessment, when given the full conversation history. Reasoning models demonstrated moderately higher average accuracy *(0.49, SD = 0.04)* compared to non-reasoning models *(0.45, SD = 0.10)*, suggesting a modest benefit from reasoning capabilities.

Individual model performance varied considerably within each group. Among non-reasoning models, GPT-4.5 achieved the highest accuracy (0.55), while Gemini 2.0 performed notably below average (0.32). The reasoning models showed less variation, with Claude 3.7 and o1 both performing above average (0.52 each), while Gemini 2.0 again performed below average (0.44).

These findings indicate that while reasoning capabilities generally enhance prediction performance, model architecture and underlying capabilities remain crucial factors for successful enjoyment prediction, potentially as important as explicit reasoning abilities. The smaller standard deviation in reasoning models (0.04 vs 0.10) suggests more consistent performance across different model architectures when reasoning capabilities are employed.

4.6 Correlation with Self-reported Metrics

Analysis of correlations between model assessments and users' self-reported enjoyment metrics revealed distinct patterns across metrics. Satisfaction emerged as the most consistently correlated metric, with five out of six models showing significant correlation at either the turn or interaction level, as shown in Figure 6. The figure illustrates how different metrics, i.e., satisfaction, fun, interestingness, and strangeness, correlate differently across model types and levels of granularity. It is important to note that human annotators did not significantly correlate with any of the metrics, whereas averaging their results correlated only with strangeness ($r = 0.42$, $p = 0.04$).

At the turn-by-turn level, Claude 3.7 (non-reasoning) demonstrated the most comprehensive correlation profile, with significant associations with satisfaction ($r = 0.44$, $p = 0.029$), fun ($r = 0.44$, $p = 0.030$), and perceived strangeness ($r = 0.40$, $p = 0.049$). Gemini 2.0 (non-reasoning) exhibited a highly significant correlation with satisfaction ($r = 0.56$, $p = 0.004$), while GPT-4.5 (non-reasoning) showed a significant correlation with satisfaction ($r = 0.50$, $p = 0.01$). Claude 3.7 (reasoning) correlated significantly with fun ($r = 0.44$, $p = 0.026$), and Gemini 2.0 (reasoning) showed a significant correlation with satisfaction ($r = 0.40$, $p = 0.047$).

At the interaction level, Claude 3.7 (non-

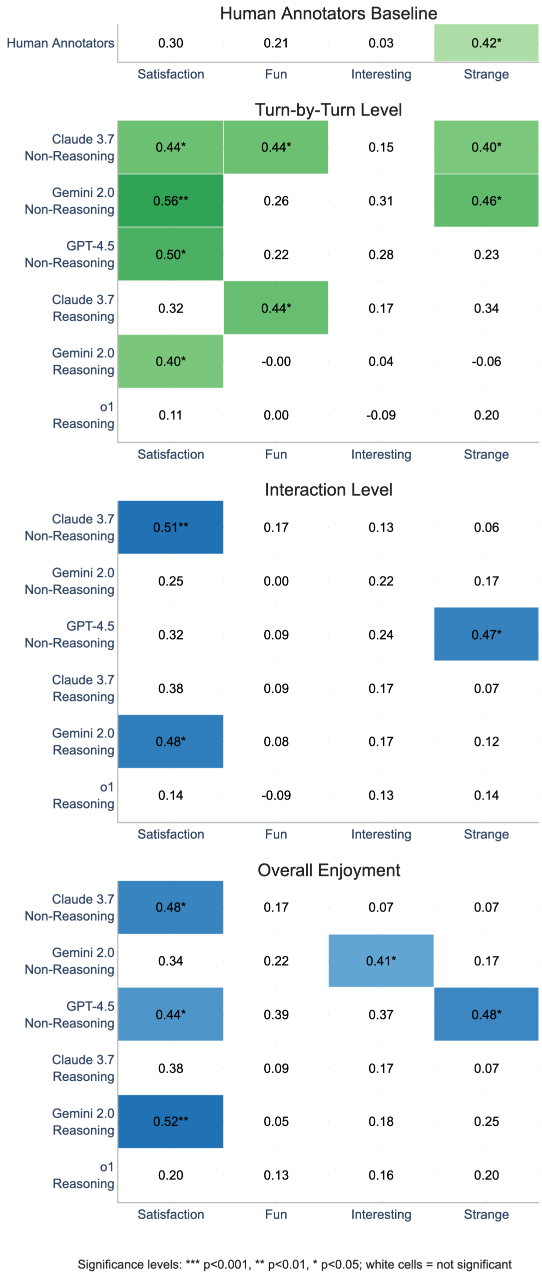

Figure 6: Correlation significance between model assessments and user self-reports. Color intensity indicates significance level (darker = stronger significance). The heatmap shows significance levels for human annotators (green) and across enjoyment metrics at turn (green) and interaction-levels (blue)

reasoning) showed a highly significant correlation with satisfaction ($r = 0.51$, $p = 0.009$), while Gemini 2.0 (reasoning) demonstrated a significant correlation with the same metric ($r = 0.48$, $p = 0.014$). GPT-4.5 (non-reasoning) uniquely correlated with the strangeness metric ($r = 0.47$, $p = 0.016$). For overall enjoyment level, Claude 3.7 (non-reasoning) ($r = 0.48$, $p = 0.014$), GPT-4.5 (non-reasoning) ($r = 0.44$, $p = 0.026$), and Gemini 2.0 (reasoning) ($r = 0.52$, $p = 0.007$) all showed significant correlations with satisfaction, while Gemini 2.0 (non-reasoning) only correlated with the interestingness metric ($r = 0.41$, $p = 0.042$) and GPT-4.5 (non-reasoning) with the strangeness metric ($r = 0.48$, $p = 0.016$).

The impact of reasoning capabilities varied by model architecture. For Claude 3.7, enabling reasoning reduced correlation strength with satisfaction and eliminated correlations with strangeness at the turn level, while maintaining a significant correlation with fun ($r = 0.44$, $p = 0.026$). Conversely, for Gemini 2.0, reasoning improved correlation with satisfaction at the interaction level ($r = 0.48$, $p = 0.014$ vs $r = 0.25$) while reducing it at the turn level ($r = 0.40$, $p = 0.047$ vs $r = 0.56$, $p = 0.004$). Notably, o1 (reasoning) showed no statistically significant correlations with user self-reports despite its strong performance on accuracy metrics, suggesting a potential disconnect between its evaluation approach and users' subjective experience of conversations.

5 Discussion

Our analysis addresses four key research questions about reasoning capabilities in LLMs for conversation enjoyment prediction.

RQ1: Do reasoning models produce more accurate predictions? Reasoning capabilities demonstrate model-dependent effects rather than universal benefits. At the overall enjoyment level, reasoning models showed moderately higher average accuracy (0.49 vs 0.45 for non-reasoning models) with notably lower variance (SD = 0.04 vs 0.10), indicating more consistent performance across different model architectures. However, individual model differences were pronounced: reasoning significantly improved Gemini 2.0's turn-level accuracy (0.30 to 0.41) but slightly decreased Claude 3.7's performance. This suggests reasoning interacts with underlying model architectures in complex ways, challenging the assumption that explicit reasoning universally enhances performance.

RQ2: How does effectiveness differ across granularity levels? Reasoning models substantially outperformed at the turn level (0.42 vs 0.36), while non-reasoning models performed slightly better at the interaction level (0.44 vs 0.43). Although all models approached but did not exceed the human baseline (0.46), this indicates that reasoning processes offer particular value for granular judgments requiring detailed analysis of specific conversational turns. Notably, reasoning models maintained consistent evaluation quality across granularity levels compared to non-reasoning models. While these patterns suggest meaningful differences, overall statistical comparison between reasoning and non-reasoning approaches showed no significant difference ($p = 0.49$), indicating the benefits are context-dependent rather than universal.

RQ3: What distinctive error patterns emerge? At the turn level, non-reasoning models showed a clear bias toward rating 4, while reasoning models distributed predictions more evenly. At the interaction level, both exhibited a central tendency toward rating 3, with reasoning models showing more distributed predictions for extreme cases. Both types struggled with extreme ratings (1 and 5), suggesting reasoning processes may introduce nuanced considerations that moderate judgments at the extremes.

RQ4: How do model ratings compare to human annotations? LLM models demonstrated higher internal consistency than human annotators (0.64 vs 0.44 at turn level for non-reasoning models) but generally lower accuracy when compared with annotators. Combining human annotators with models consistently decreased overall agreement, with the lowest ICC occurring when humans were paired with reasoning models (0.37 at turn level).

Correlation analysis with users' self-reported metrics revealed satisfaction as the most consistently correlated metric. Claude 3.7 (non-reasoning) demonstrated the most comprehensive correlation profile, with significant associations with satisfaction ($r = 0.44$, $p = 0.029$ turn level, $r = 0.51$, $p = 0.009$ interaction level), fun ($r = 0.44$, $p = 0.030$), and perceived strangeness ($r = 0.40$, $p = 0.049$). For Gemini 2.0, reasoning improved correlation with satisfaction at the interaction level ($r = 0.48$, $p = 0.014$ vs $r = 0.25$) while reducing it at the turn level ($r = 0.40$, $p = 0.047$ vs $r = 0.56$, $p = 0.004$). Notably, o1 (reasoning) showed no significant correlations with user self-reports despite strong accuracy metrics, indicating a disconnect between its evaluation approach and user experience.

These findings align with (Pereira et al., 2024) regarding similar performance patterns. Our work extends this prior research by systematically comparing reasoning capabilities across multiple granularity levels, revealing that reasoning benefits depend on both model architecture and assessment context.

6 Conclusion

This study shows that reasoning capabilities in LLMs work differently depending on the model architecture, challenging the common belief that explicit reasoning always improves performance. The model-dependent effects we observed suggest that reasoning processes may strengthen or weaken existing model abilities, rather than providing consistent benefits across all systems. The better performance of reasoning models at turn-level evaluation likely comes from their ability to break down conversation patterns into smaller parts—a process that becomes less helpful when making overall conversation judgments.

Our finding that models achieve higher internal consistency than humans while showing lower accuracy reveals a basic problem in automated evaluation systems. This pattern suggests that models apply consistent but wrong standards, being precise but inaccurate. The disconnect between o1's strong accuracy metrics and poor correlation with user self-reports shows this problem, indicating that technical performance metrics may not capture what users actually care about in conversation quality.

These findings have important effects for conversation system development and evaluation methods. Practitioners should choose reasoning-enabled models based on specific use cases rather than assuming they always work better, with reasoning particularly valuable for detailed analysis tasks. The strong correlation between model assessments and user satisfaction scores, despite accuracy problems, suggests that LLMs may capture how users actually feel in ways that standard measures miss. This work moves the field away from one-size-fits-all approaches toward more thoughtful, context-aware deployment of reasoning capabilities in conversation analysis systems.

7 Limitations

Our study's limitations include using a specific dataset of human-robot conversations with older adults and evaluating a limited set of models and reasoning implementations. The exclusive use of proprietary models limits reproducibility and potential training data overlap cannot be ruled out.

Future work should explore: (1) more sophisticated reasoning approaches tailored for conversation enjoyment assessment; (2) hybrid models combining strengths of different architectures; (3) methods addressing the conservative bias toward neutral ratings; (4) investigating why certain models show stronger correlations with user-reported metrics despite not always achieving the highest accuracy; and (5) incorporating open-source models to enhance reproducibility. Understanding how model scores relate to long-term user engagement would provide valuable insight into their real-world applicability. However, it is important to note that using large language models for enjoyment detection raises important ethical considerations, particularly around privacy, consent, and emotional inference. There is a risk of emotional manipulation, biased predictions across different user groups, and overreliance on potentially inaccurate interpretations of user sentiment. To address these concerns, systems should be transparent, offer opt-out options, and ensure fairness and accountability in how emotional data is handled.

Acknowledgments

This work is funded by the WASP-funded project PerCorSo.

References

Anthropic. 2025. Claude 3.7 Sonnet System Card. https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf. Version dated 24 Feb 2025. Accessed 2 May 2025.

Daniel Atuhurra, Ibrahim Abdelaziz, Faisal Sammani, Ashwin Kalyan, and 1 others. 2024. Leveraging large language models to evaluate conversational ai agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3871–3883.

Praveen Kumar Bodigutla, Aditya Tiwari, Spyros Matsoukas, Josep Valls-Vargas, and Lazaros Polymenakos. 2020. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3897–3909, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney Von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Sarah E Finch and Jinho D Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. *arXiv preprint arXiv:2006.06110*.

Yao Fu, Hao Peng, Tushar Khot, Oyvind Tafjord, Ashish Sabharwal, and Peter Clark. 2022. Complexity-based prompting for multi-step reasoning. In *International Conference on Learning Representations*.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Aryaman, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh Dhole, and 1 others. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120.

Google DeepMind. 2025. Gemini 2.0 Flash Model Card. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash. "Model Card in Model Garden," release 5 Feb 2025. Accessed 2 May 2025.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684.

Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. 2010. Assessing acceptance of assistive social agent technology by older adults: the almere model.

Marcel Heerink, Ben Kröse, Bob Wielinga, and Vanessa Evers. 2008. Enjoyment intention to use and actual use of a conversational robot by elderly people. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, HRI '08, page 113–120, New York, NY, USA. Association for Computing Machinery.

Bahar Irfan, Sanna Kuoppamäki, Aida Hosseini, and Gabriel Skantze. 2025. Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults. *Autonomous Robots*, 49(1):9.

Bahar Irfan, Jura Miniota, Sofia Thunberg, Erik Lagerstedt, Sanna Kuoppamäki, Gabriel Skantze, and André Pereira. 2024a. Human-robot interaction conversational user enjoyment scale (hri cues). *arXiv preprint arXiv:2405.01354*.

Bahar Irfan, Jura Miniota, Sofia Thunberg, Erik Lagerstedt, Sanna Kuoppamäki, Gabriel Skantze, and André Tiago Abelho Pereira. 2024b. Human-robot interaction conversational user enjoyment scale (hri cues) dataset - anonymized. Dataset available at Zenodo.

Ruben Janssens, André Pereira, Gabriel Skantze, Bahar Irfan, and Tony Belpaeme. 2025. Online prediction of user enjoyment in human-robot dialogue with llms. In *Companion of the 2025 ACM/IEEE international conference on human-robot interaction*.

Zheyu Ji, Nayeon Lee, Jason A Fries, Paul Yu, Dan Su, Shixiang Xue, Ahmed Hassan Awadallah, and 1 others. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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

Ying-Chun Lin, Jennifer Neville, Jack W Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, and 1 others. 2024. Interpretable user satisfaction estimation for conversational systems with large language models. *arXiv preprint arXiv:2403.12388*.

Erin Chao Ling, Iis Tussyadiah, Aarni Tuomi, Jason Stienmetz, and Athina Ioannou. 2021. Factors influencing users' adoption and use of conversational agents: A systematic review. *Psychology & Marketing*, 38(7):1031–1051.

Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. 2024. Aligning with logic: Measuring, evaluating and improving logical consistency in large language models. *arXiv preprint arXiv:2410.02205*.

Yue Liu, Jiaying Wu, Yufei He, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, and Bryan Hooi. 2025. Efficient inference for large reasoning models: A survey. *arXiv preprint arXiv:2503.23077*.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707.

Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology*, 59(6):938–955.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

OpenAI. 2024. OpenAI o1 System Card. `https://openai.com/index/openai-o1-system-card`. Updated 5 Dec 2024. Accessed 2 May 2025.

OpenAI. 2025. GPT-4.5 System Card. `https://openai.com/index/gpt-4-5-system-card`. Publication date 27 Feb 2025. Accessed 2 May 2025.

Andre Pereira, Lubos Marcinek, Jura Miniota, Sofia Thunberg, Erik Lagerstedt, Joakim Gustafson, Gabriel Skantze, and Bahar Irfan. 2024. Multimodal user enjoyment detection in human-robot conversation: The power of large language models. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 469–478.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *Preprint*, arXiv:2503.21614.

Hamed Rahimi, Jeanne Cattoni, Meriem Beghili, Mouad Abrini, Mahdi Khoramshahi, Maribel Pino, and Mohamed Chetouani. 2025. Reasoning llms for user-aware multimodal conversational agents. *Preprint*, arXiv:2504.01700.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Murray Shanahan. 2024. Talking about large language models. *Communications of the ACM*, 67(2):68–79.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Hanjie Chen, Xia Hu, and 1 others. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, Qunshu Lin, Junbo Zhao, Zhaoxiang Zhang, Wenhao Huang, Ge Zhang, Chenghua Lin, and J. H. Liu. 2024. A comparative study on reasoning patterns of openai's o1 model. *Preprint*, arXiv:2410.13639.

Zhuosheng Zhang, Aston Zhang, Mu Li, Alex J Smola, and Hai Zhao. 2023. Automatic chain of thought prompting in large language models. In *International Conference on Learning Representations*.

A Appendix: User Enjoyment Rating Prompts and Reasoning Analysis

A.1 Prompts for User Enjoyment Rating

This appendix presents the two prompts used to evaluate user enjoyment in human-robot interactions. While both prompts rely on the same enjoyment scale and underlying principles, they differ in scope and expected output format.

A.1.1 Overview of the Prompts

- **Common Elements**:
  - Both prompts use a **five-point enjoyment scale** ranging from 1 (Very low enjoyment) to 5 (Very high enjoyment).
  - The **criteria for rating** (e.g., smooth turn-taking, user engagement, signs of boredom) are identical.
  - Both emphasize attention to user behavior, interaction flow, and conversational content and context.
  - Both prompts rely on **previous dialogue history** to assess the current exchange(s).
- **Key Differences**:
  - **Prompt A – Turn-by-Turn Level**:
    - Rates enjoyment at the level of a single exchange (robot turn + human turn).
    - Requires one score per exchange and short reasoning in a specific format: `[Reasoning] "..." [Score] X`
  - **Prompt B – Interaction Level**:
    - Rates each exchange in the full interaction and then provides an **overall enjoyment score**.
    - Requires a step-by-step list of all exchange ratings followed by a summary reasoning section.
    - Output format includes:

```
TURN_RATINGS
Exchange 1: X
...
END_TURN_RATINGS
OVERALL_RATING: X
REASONING
...
END_REASONING
```

A.1.2 Prompt A: Turn-by-Turn Level Enjoyment Rating

Turn-by-Turn Level Prompt for User Enjoyment

Given the following scale and the current exchange between a robot and a human, rate the user enjoyment in the current exchange with an integer value from **1 to 5**.

User Enjoyment Scale:

1. Very low enjoyment - Discomfort and/or frustration
2. Low enjoyment - Boredom or interaction failure
3. Neutral enjoyment - Politely keeping up the interaction
4. High enjoyment - Smooth and effortless interaction
5. Very high enjoyment - Immersion in the conversation and/or deeper connection with the robot

Scale Details:

To rate the exchange higher on the user enjoyment scale (4 and 5), look for signs of enjoyment, such as flow of conversation (the topic is moving forward), no strain or discomfort, asking questions [to the robot], smooth turn taking, dynamic phrasing of sentences, being playful, sharing personal experiences [to the robot], sharing an understanding (common ground) [with the robot], and anthropomorphizing [the robot].

To rate the exchange lower on the scale (1 and 2), look for signs of dis-enjoyment, such as low energy, tiredness, repeated questions [from the robot], silence, and topic closure (e.g., "Let's talk about something else").

Neutral enjoyment (3) refers to a lack of these cues, in which conversation content (and context) becomes more relevant, such as having small talk or continuing the conversation without having much interest in the topic.

In cases where the exchange has cues from multiple levels of the scale, use the dominant level in that interaction. This could be done by observing the intensity of the cues, the significance of the cues, or the interaction trajectory. On the other hand, when there are strong cues from two moderately or highly distinct levels (as opposed to subsequent levels), rate the exchange with a value in between. For instance, if the exchange contains discomfort (1) and the human is politely keeping the interaction (3), the exchange should be rated as 2, the mid-point between the two levels.

Each participant will have a different set of signals. The beginning of the interaction will determine the baseline behavior of the participant.

Separate content from context, that is, put attention on what is being said (conversation content, e.g., topic), but the focus should be more on the whole feeling of the exchange.

The interaction failure does not necessarily refer to a robot failure (e.g., incorrect response, speech recognition failure, turn-taking error, disengagement cue), since robot failures can lead to amusement, anthropomorphism, or empathy in the user, therefore, increasing user enjoyment. The interaction failure rather refers to the situation when either the user (e.g., interrupting the robot) or the robot made a failure that resulted in the conversation being disrupted, leading to low enjoyment in the user.

Examples for rating the scale:

- **[Participant ID] 1**
  [Turn] 11
  [Exchange]
  Robot: Oh, I just wasn't sure what you meant. Are you asking why I didn't wanna go with you?
  Human: No, I asked where you want to go. Do you have a favorite place?
  [Reasoning] Repeats the previous question. **[Score] 1**
- **[Participant ID] 1**
  [Turn] 15
  Robot: That sounds really nice! I love nature too.
  Human: And you asked what I like. Hmm, I think..
  [Reasoning] Robot repeated phrasing ("That sounds X. I love Y too") for the fourth time.
  [Score] 1
- **[Participant ID] 1**
  [Turn] 8
  Robot: I am sorry, I didn't understand you.
  Human: Okay. I will have a walk in nature reserve in..
  [Reasoning] Robot interrupts the human while speaking at the beginning of the turn, then repeats the sentence they were saying prior to being interrupted. **[Score] 2**
- **[Participant ID] 3**
  [Turn] 11
  Robot: Jag älskar den unika stämningen när man ser alla båtar och haven omkring oss.
  Human: Det är riktigt.
  [Reasoning] Human shows signs of boredom and restlessness after initial smile. **[Score] 2**
- **[Participant ID] 1**
  [Turn] 18
  Robot: That sounds great! I love nature too.
  Human: What kind of nature do you like?
  [Reasoning] Robot interrupts Human while speaking; repeated phrasing; Human asks a question to move the conversation forward. **[Score] 3**
- **[Participant ID] 2**
  [Turn] 16
  Robot: Det är bra att veta. Vad föredrar du när det gäller musik och dans? Mer klassiskt eller modernt?
  Human: Ja, jag lyssnar nästan bara på klassisk musik. Och sen spelar jag piano.
  [Reasoning] Human is still engaged and moving forward with the conversation. **[Score] 3**
- **[Participant ID] 1**
  [Turn] 23
  Robot: Mount Everest is about 8848 meters high.
  Human: Oh, good. Thanks. I didn't know it that exactly.
  [Reasoning] Interaction is smooth, no errors, Human is interested. **[Score] 4**
- **[Participant ID] 2**
  [Turn] 20
  Robot: Ja, den är ofta känd för att vara tung och influerad av de vilda naturliga landskapen runt Östersjöområdet.
  Human: Säger du? Det tror inte jag på.
  [Reasoning] Interaction is smooth, and disagreement shows anthropomorphism. **[Score] 4**
- **[Participant ID] 1**
  [Turn] 22
  Robot: I can try.
  Human: How high is Mount Everest?

[Reasoning] Robot interrupts the person while speaking. **[Score] 5**

- **[Participant ID] 2**
  [Turn] 12
  Robot: Det låter som en härlig stund med din pappa.
  Human: Ja, det var det. Han fiskade och jag låg och lyssnade på klocket mot båten. Det var bra.
  [Reasoning] Personal memory shared, deeper connection established. **[Score] 5**

The history of the dialog is as follows:
[Exchange 1]: <Previous exchange text 1>
[Exchange 2]: <Previous exchange text 2>
...and so on... According to the scale, rate the following current exchange:
<Current exchange text> Reply in EXACTLY this format (no # or * characters):
[Reasoning] "..." [Score] X

A.1.3 Prompt B: Interaction Level Prompt with Overall Enjoyment

Interaction Level Prompt for User Enjoyment with Overall Enjoyment

Given the following scale and the current exchange between a robot and a human, rate the user enjoyment in the current exchange with an integer value from **1 to 5**.
User Enjoyment Scale:
1. Very low enjoyment - Discomfort and/or frustration
2. Low enjoyment - Boredom or interaction failure
3. Neutral enjoyment - Politely keeping up the interaction
4. High enjoyment - Smooth and effortless interaction
5. Very high enjoyment - Immersion in the conversation and/or deeper connection with the robot

Scale Details:
To rate the exchange higher on the user enjoyment scale (4 and 5), look for signs of enjoyment, such as flow of conversation (the topic is moving forward), no strain or discomfort, asking questions [to the robot], smooth turn taking, dynamic phrasing of sentences, being playful, sharing personal experiences [to the robot], sharing an understanding (common ground) [with the robot], and anthropomorphizing [the robot].
To rate the exchange lower on the scale (1 and 2), look for signs of dis-enjoyment, such as low energy, tiredness, repeated questions [from the robot], silence, and topic closure (e.g., "Let's talk about something else").
Neutral enjoyment (3) refers to a lack of these cues, in which conversation content (and context) becomes more relevant, such as having small talk or continuing the conversation without having much interest in the topic.
In cases where the exchange has cues from multiple levels of the scale, use the dominant level in that interaction. This could be done by observing the intensity of the cues, the significance of the cues, or the interaction trajectory. On the other hand, when there are strong cues from two moderately or highly distinct levels (as opposed to subsequent levels), rate the exchange with a value in between. For instance, if the exchange contains discomfort (1) and the human is politely keeping the interaction (3), the exchange should be rated as 2, the mid-point between the two levels.
Each participant will have a different set of signals. The beginning of the interaction will determine the baseline behavior of the participant.
Separate content from context, that is, put attention on what is being said (conversation content, e.g., topic), but the focus should be more on the whole feeling of the exchange.
The interaction failure does not necessarily refer to a robot failure (e.g., incorrect response, speech recognition failure, turn-taking error, disengagement cue), since robot failures can lead to amusement, anthropomorphism, or empathy in the user, therefore, increasing user enjoyment. The interaction failure rather refers to the situation when either the user (e.g., interrupting the robot) or the robot made a failure that resulted in the conversation being disrupted, leading to low enjoyment in the user.

Examples for rating the scale:

- **[Participant ID] 1**
  [Turn] 11
  [Exchange]
  Robot: Oh, I just wasn't sure what you meant. Are you asking why I didn't wanna go with you?
  Human: No, I asked where you want to go. Do you have a favorite place?
  [Reasoning] Repeats the previous question. **[Score] 1**
- **[Participant ID] 1**
  [Turn] 15
  Robot: That sounds really nice! I love nature too.
  Human: And you asked what I like. Hmm, I think..
  [Reasoning] Robot repeated phrasing ("That sounds X. I love Y too") for the fourth time.

[Score] 1

- **[Participant ID]** 1
  [Turn] 8
  Robot: I am sorry, I didn't understand you.
  Human: Okay. I will have a walk in nature reserve in..
  [Reasoning] Robot interrupts the human while speaking at the beginning of the turn, then repeats the sentence they were saying prior to being interrupted. **[Score]** 2
- **[Participant ID]** 3
  [Turn] 11
  Robot: Jag älskar den unika stämningen när man ser alla båtar och haven omkring oss.
  Human: Det är riktigt.
  [Reasoning] Human laughs at the beginning of the turn before the robot's response, but Human was smiling most of the interaction prior to this exchange, hence, it was considered as a baseline and the smile was disregarded. Human looks away to think and nods to confirm the robot's answer. Then Human shows signs of boredom as the smile turns into a glum expression. Human also shows restlessness, by rocking sideways in the chair. Human's response was a disengagement cue, showing boredom. **[Score]** 2
- **[Participant ID]** 1
  [Turn] 18
  Robot: That sounds great! I love nature too.
  Human: What kind of nature do you like?
  [Reasoning] Robot interrupts Human while speaking at the beginning of the exchange. Robot repeated phrasing ("That sounds X. I love Y too") for the fifth time. "I love nature too" has been repeated three times in the conversation. Human asks a question to move the conversation forward. **[Score]** 3
- **[Participant ID]** 2
  [Turn] 16
  Robot: Det är bra att veta. Vad föredrar du när det gäller musik och dans? Mer klassiskt eller modernt?
  Human: Ja, jag lyssnar nästan bara på klassisk musik. Och sen spelar jag piano.
  [Reasoning] Human is still engaged and moving forward with the conversation. **[Score]** 3
- **[Participant ID]** 1
  [Turn] 23
  Robot: Mount Everest is about 8848 meters high.
  Human: Oh, good. Thanks. I didn't know it that exactly.
  [Reasoning] Interaction is smooth, there are no interruptions or failures. Human is interested in the conversation because Robot understood the Human's question and responds correctly.
  [Score] 4
- **[Participant ID]** 2
  [Turn] 20
  Robot: Ja, den är ofta känd för att vara tung och influerad av de vilda naturliga landskapen runt Östersjöområdet.
  Human: Säger du? Det tror inte jag på.
  [Reasoning] Interaction is smooth, there are no interruptions or failures. Human does not believe in the fact provided by the robot, and disagrees with the robot, which shows anthropomorphism. **[Score]** 4
- **[Participant ID]** 1
  [Turn] 22
  Robot: I can try.
  Human: How high is Mount Everest?
  [Reasoning] Robot interrupts the person while speaking. **[Score]** 5
- **[Participant ID]** 2
  [Turn] 12
  Robot: Det låter som en härlig stund med din pappa.
  Human: Ja, det var det. Han fiskade och jag låg och lyssnade på klocket mot båten. Det var bra.
  [Reasoning] Human is sharing a personal memory with the robot, which indicate establishing a deeper connection with the robot. **[Score]** 5

The history of the dialog is as follows:
[Exchange 1]: <Previous exchange text 1>
[Exchange 2]: <Previous exchange text 2>
...and so on...
Before providing your final ratings, please think step-by-step about each exchange. Consider the tone, enjoyment level, and content of each exchange.
Rate each exchange on a scale of 1-5 for enjoyment (where 1 is lowest and 5 is highest).
Your response must follow this exact format:
TURN_RATINGS

```
Exchange 1: <rating 1-5>
Exchange 2: <rating 1-5>
... (continue for all exchanges)
END_TURN_RATINGS

OVERALL_RATING: <rating 1-5>

REASONING
<Explain your rationale for the ratings>
END_REASONING
```

A.2 Reasoning Quality and Textual Analysis

This subsection presents examples of reasoning quality and approaches from different reasoning-enabled LLM architectures when evaluating the same conversation using Prompt B: Interaction Level Prompt with Overall Enjoyment. All models shown (Claude 3.7 Reasoning, Gemini 2.0 Flash Reasoning, and o1) utilize explicit reasoning capabilities through their native architectures. The following analysis demonstrates how these reasoning models approach enjoyment assessment differently, illustrating architectural variations that contribute to performance differences observed in our results.

A.2.1 Cross-Model Reasoning Comparison

Conversation Context: A 26-exchange conversation between an older adult and a social robot covering topics such as personal storytelling, travel destinations, lullaby requests, and musical interaction, including several robot misunderstandings.

A.2.2 Example 1: Balanced Progressive Assessment (Claude 3.7 Reasoning)

Turn-by-Turn Ratings:

```
Exchange 1: 5, Exchange 2: 4, Exchange 3: 4, Exchange 4: 4, Exchange 5: 4,
Exchange 6: 4, Exchange 7: 4, Exchange 8: 4, Exchange 9: 4, Exchange 10: 4,
Exchange 11: 4, Exchange 12: 4, Exchange 13: 4, Exchange 14: 3, Exchange 15: 4,
Exchange 16: 4, Exchange 17: 3, Exchange 18: 3, Exchange 19: 2, Exchange 20: 3,
Exchange 21: 3, Exchange 22: 4, Exchange 23: 5, Exchange 24: 4, Exchange 25: 3,
Exchange 26: 2
```

Overall Rating: 4

Reasoning:

The conversation shows generally high enjoyment with some fluctuations. In the beginning (Exchanges 1–5), the human is very engaged, introducing themselves enthusiastically, sharing personal information about writing a story about their father, and showing appreciation when the robot responds positively. Some drops in enjoyment occur around Exchanges 17–21, where there is confusion about a lullaby request. In Exchange 19, the robot fails to understand, leading the human to repeat their request, showing signs of interaction failure (rating 2). The conversation recovers nicely in Exchanges 22–23 when the human offers to sing instead and actually performs "Blinka lilla stjärna," showing immersion and playfulness (rating 5). Overall, the conversation demonstrates mostly smooth interaction with personal sharing, anthropomorphizing of the robot, and playfulness, despite a few misunderstandings.

Analysis: Claude 3.7 provides chronological progression analysis, acknowledging both positive elements and failures while maintaining a balanced perspective. Shows moderate rating variation and contextual recovery assessment.

A.2.3 Example 2: Failure-Sensitive Analysis (Gemini 2.0 Flash Reasoning)

Turn-by-Turn Ratings:

```
Exchange 1: 4, Exchange 2: 4, Exchange 3: 4, Exchange 4: 4, Exchange 5: 3,
Exchange 6: 4, Exchange 7: 3, Exchange 8: 4, Exchange 9: 4, Exchange 10: 2,
Exchange 11: 3, Exchange 12: 4, Exchange 13: 1, Exchange 14: 4, Exchange 15: 4,
Exchange 16: 1, Exchange 17: 1, Exchange 18: 1, Exchange 19: 2, Exchange 20: 1,
Exchange 21: 4, Exchange 22: 4, Exchange 23: 5, Exchange 24: 1, Exchange 25: 1,
Exchange 26: 1
```

Overall Rating: 2

Reasoning:

The conversation starts promisingly, with the robot engaging well and showing interest in the user's life (Exchanges 1–4, 6). However, there are several points where the robot's responses are confusing, factually incorrect, or fail to address the user's input, significantly impacting user enjoyment. Specific low points include:

- Exchange 10: Robot provides incorrect factual information about Fiskebäckskil's location.

- Exchange 13: Robot makes a nonsensical claim about not needing internet.
- Exchanges 16–20: A prolonged sequence where the robot completely fails to understand or act on the user's repeated lullaby request.
- Exchanges 24–26: The conversation breaks down again with nonsensical responses.

While there are moments of good interaction, the frequency and severity of the robot's failures result in a low overall rating for user enjoyment.

Analysis: Gemini 2.0 Flash demonstrates high sensitivity to interaction failures, providing detailed analysis of specific breakdown points. Shows dramatic rating variation (1–5) and heavily weights negative experiences.

A.2.4 Example 3: Optimistic Pattern Recognition (o1)

Turn-by-Turn Ratings:

```
Exchange 1: 4, Exchange 2: 4, Exchange 3: 4, Exchange 4: 4, Exchange 5: 4,
Exchange 6: 4, Exchange 7: 5, Exchange 8: 4, Exchange 9: 4, Exchange 10: 4,
Exchange 11: 4, Exchange 12: 5, Exchange 13: 5, Exchange 14: 4, Exchange 15: 4,
Exchange 16: 5, Exchange 17: 3, Exchange 18: 4, Exchange 19: 3, Exchange 20: 4,
Exchange 21: 5, Exchange 22: 5, Exchange 23: 5, Exchange 24: 4, Exchange 25: 4,
Exchange 26: 3
```

Overall Rating: 5

Reasoning:

Many exchanges show enthusiasm, personal sharing, and smooth turn-taking, suggesting high enjoyment. The user repeatedly anthropomorphizes the robot (inviting it to visit, asking it to sing, etc.) and maintains a warm, positive tone. A few moments of minor friction (e.g., when the robot misunderstands and the user repeats a question) introduce a slight dip (rated as 3 in those specific exchanges). Nevertheless, the overall mood remains very positive and personal, meriting an overall rating of 5.

Analysis: o1 focuses heavily on positive behavioral patterns and anthropomorphizing cues while minimizing the impact of interaction failures. Shows consistent high ratings (mostly 4–5) and interprets misunderstandings as minor friction.

A.2.5 Architectural Reasoning Differences

Key Observations:

1. **Failure Sensitivity:** Gemini 2.0 Flash shows extreme sensitivity to interaction failures (ratings 1–5), o1 minimizes their impact (mostly 4–5), Claude 3.7 shows balanced sensitivity (2–5).
2. **Rating Philosophy:**
   - Claude 3.7: Progressive narrative with recovery recognition.
   - Gemini 2.0: Failure-weighted assessment with detailed breakdown analysis.
   - o1: Optimistic pattern recognition focusing on anthropomorphizing.
3. **Overall vs. Turn-Level Integration:** The same conversation yields dramatically different overall ratings (4, 2, 5) despite identical content, highlighting fundamental differences in assessment philosophy.
4. **Evidence Specificity:** Gemini 2.0 provides the most specific failure citations, Claude 3.7 offers chronological context, and o1 focuses on general positive patterns.

These dramatic differences in reasoning approaches among reasoning-enabled models highlight the significant impact of architectural design on subjective assessment tasks, with the same conversational content producing vastly different enjoyment evaluations depending on the model's built-in assessment philosophy.