Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Automating Dialogue Evaluation: Large Language Mode versus Human Judgement

Ebubechukwu Ike, Johane Takeuchi, Antonello Ceravola, Frank Joublin

2025

Preprint:

This is a post-peer-review, pre-copyedit version of an article published in HCI International 2025. The final authenticated version is available online at: https://doi.org/10.1007/978-3-031-93415-5_21

Automating Dialogue Evaluation: LLMs vs Human Judgment

Ebubechukwu Ike[1,4][0009−0008−7228−1434], Johane Takeuchi[2][0000−0002−4010−4491], Frank Joublin[1][0000−0002−4421−1737], Antonello Ceravola[1][0000−0002−1075−459X], and Marc Tanti[3][0000−0002−7156−0596]

[1] Honda Research Institute Europe, Offenbach am Main, Germany
{frank.joublin, antonello.ceravola}@honda-ri.de
[2] Honda Research Institute Japan, Wako, Japan
johane.takeuchi@honda-ri.jp
[3] University of Malta, Msida, Malta
marc.tanti@um.edu.mt
[4] University of Bielefeld, Germany
ebubechukwu.ike@uni-bielefeld.de

Abstract. As dialogue systems and chatbots become more common in daily life, efficient and accurate evaluation methods are crucial. This study compares human and AI assessments across various dialogue scenarios, focusing on seven key performance indicators (KPIs): Coherence, Innovation, Concreteness, Goal Contribution, Commonsense Contradiction, Incorrect Fact, and Redundancy. Using the GPT-4o API, we generated diverse conversation datasets and conducted a two-part analysis. Experiment 1 evaluated multi-party conversations on Coherence, Innovation, Concreteness, and Goal Contribution, finding that GPT models closely match human judgments. Experiment 2 focused on dyadic dialogues, assessing Commonsense Contradiction, Incorrect Fact, and Redundancy, showing GPT-4o excels in maintaining factual accuracy and commonsense reasoning but struggles with redundancy and self-contradiction. These findings highlight GPT models' potential to replicate human evaluation in dialogue systems while identifying areas for improvement.

Keywords: Dialog Systems · Multi-party Dialogues · Human-AI Interaction · Dialogue Evaluation · Key Performance Indicators (KPI) · GPT-4o

1 Introduction

Communication is a fundamental aspect of life, serving as the primary mechanism for interaction among living organisms. To improve the quality of these dialogues, evaluation methods were introduced. Initially relying on resources like dictionaries to assess contextual appropriateness [45], and with the advent of linguistics, more sophisticated metrics emerged to systematically evaluate dialogues [14]. While human evaluators have been effective, their involvement is resource-intensive,

costly, and can be prone to biases and errors [15]. Tools like Microsoft Word were also developed to assist by automatically detecting spelling and grammatical errors, streamlining the human review process [24, 48] and helping improve quality in communication overall.

As technology advanced, models were developed to evaluate dialogues more effectively, analyzing word usage, sentence structures, or even sentiment [40]. These models, particularly those automatable, found applications in social media for tasks like detecting inappropriate content ([10], [50]). However, they were limited in replicating human understanding and nuance [4]. The introduction of AI, particularly chatbots [52], revolutionized customer interaction. Today, around 80% of companies use chatbots for customer interaction [19][5]. Human expertise combined with machine learning therefore ensured accuracy and quality in chatbot responses, though primarily focused on surface-level metrics like word accuracy and sentence comprehension [26].

Advancements in conversational AI then led to robots capable of both task-oriented and non-task-oriented dialogues, switching based on context and speech recognition outcomes [38]. This capability is crucial for robots in dynamic environments, where both functional and social interactions are essential. Integrating social dialogue enhanced a robot's influence on human decision-making, underscoring the importance of combining relational and task-oriented dialogues in human-robot interactions [34].

The advent of Generative Pre-trained Transformers (GPT), especially GPT-3, marked significant improvements in generating human-like conversations [6]. Following GPT-3, other models, like LLaMA, further advanced the field, showing comparable performance with fewer parameters and less training data [37]. These large language models (LLMs) generates conversation nearly indistinguishable from human ones. However, issues like bias and hallucinations remain, undermining trust in AI systems [4] [29].

Despite these challenges, LLMs are now widely accepted and integrated into various sectors, from education to everyday smartphone use [41][6]. As these models evolved, systems using them can now engage in dialogues with or without human intervention [21]. This has created a need for new dialogue evaluation approaches to generate more reliable and trustworthy AI dialogues.

Our aim is to contribute to automatic dialogue evaluation using the GPT-4o model, comparing its performance with human annotators. We hypothesize that if GPT-4o evaluates dialogues similarly to humans, two key outcomes will emerge: 1) reduced reliance on expensive human annotators, and 2) increased use of Multi-Agent Systems for dialogue generation, with AI-based evaluators providing comparable support to human evaluators. We focus on per-turn evaluation, which can be used to improve ongoing dialogues.

[5] Netomi
[6] JPT

Our evaluations target the semantic meaning of turns within multi-party conversations. For human evaluation, we will use user surveys with multi-score questions, automating the process with an LLM-based evaluator.

Our study will contribute to the ongoing discourse on AI reliability and the future of human-AI interaction, ensuring that dialogues remain a trusted and integral part of our communication landscape. To this end, our work focuses on two primary objectives:

1. To measure how closely GPT-4o can approximate human-level performance in multi-party conversation quality assesments.
2. To measure how closely GPT-4o can approximate human-level performance in diadic dialogue turns error classification.

2 Background

2.1 Evaluation Methods

The evaluation of dialogue systems has evolved from manual human assessments to sophisticated automated and machine learning-based methods. This section provides an overview of the historical development and current state of dialogue evaluation techniques.

1. **Human-based Evaluation:** Initially, dialogue systems were evaluated through human judgments, focusing on aspects like coherence, relevance, informativeness, and engagement [13][23][27]. Metrics such as *user satisfaction* and *task success rates* were commonly used [17]. While human evaluations are considered the gold standard due to their ability to capture nuanced dialogue qualities, they are costly, time-consuming, and prone to inter-annotator variability and bias [42][13].
2. **Rule-based and Statistical Models:** With advances in natural language processing, rule-based and statistical models emerged to automate evaluations. These methods used pre-defined rules or metrics like BLEU and ROUGE to assess the *overlap* between generated responses and reference texts [36][47]. However, these metrics often missed the full quality of dialogues, focusing too much on lexical similarities and ignoring the diversity of valid responses [30].
3. **Machine Learning Approaches:** The introduction of machine learning brought more advanced methods. Supervised models were trained on annotated datasets to predict human evaluation scores. Models like ADEM (Automatic Dialogue Evaluation Model) predict *dialogue quality scores* based on human annotations, showing improvements over rule-based metrics but still relying heavily on large, high-quality datasets [33][44].
4. **Deep Learning and Neural Networks:** Recent developments have shifted towards deep learning, particularly with models like BERT and GPT, fine-tuned for dialogue evaluation using large-scale datasets to better capture context and nuance [39][2]. Models like RUBER integrate both reference and unreferenced-based evaluation metrics, combining the strengths of both approaches [3].

Evaluation Challenges in Multi-party Dialogue: Evaluating multi-party dialogue is more complex than dyadic systems, involving intricate interactions and context management across multiple participants. Traditional metrics often fail to capture these dynamics. Expanded taxonomies and evaluation frameworks, like those proposed by [25], address these challenges, including errors specific to multi-party dialogues such as *speaker and addressee recognition errors* [35][51].

Towards Standardized Evaluation Frameworks: The field has acknowledged the need for standardized evaluation frameworks to consistently assess dialogue systems across different tasks and contexts. Initiatives like the Dialogue System Technology Challenge (DSTC) series aim to provide standardized datasets and evaluation protocols, promoting transparency and reproducibility in evaluation practices [43][8]. In summary, dialogue system evaluation has progressed from human-based methods to more automated and standardized approaches. Despite advancements, ongoing research is essential to address current challenges and improve the reliability of dialogue evaluations. This paper proposes a meta-analysis comparing LLM-based evaluation methods with human evaluation, aiming to support the transition to fully automated conversation quality assessments.

2.2 Dataset

Existing dialogue datasets cater to various topics and domains, each with its strengths and limitations. The Personal-Chat dataset [54] focuses on personal conversations, making it ideal for personalized dialogue studies but limited in scope for multi-domain applications. The MultiWOZ dataset [7] is comprehensive for task-oriented dialogues across multiple domains but is more suited for task-specific exchanges than open-domain dialogue.

Datasets like those from the TV show "Friends" [9] and OpenSubtitles [11] offer large volumes of dialogues but require significant processing to extract relevant data. DREAM [46], designed for reading comprehension, and the Entailment Bank dataset [12], focused on explanatory tasks, lack the conversational structure needed for this study. Similarly, ReDial [32], while useful for conversational recommendations, is too domain-specific for our open-domain, multi-party requirements.

Given the limitations of existing datasets, we generated a custom dataset tailored for multi-domain and multi-party dialogues, avoiding the need for extensive preprocessing. Additionally, an established benchmark dataset was used to rigorously evaluate dialogue error behaviors in a contextually relevant manner.

3 Experiments

The experiments were conducted in two parts. The first experiment involved generating a custom dataset using GPT-4o, designed for multi-party and open-domain dialogues. A total of 48 dialogues were created, each selected based on key performance indicators (KPIs) to ensure alignment with

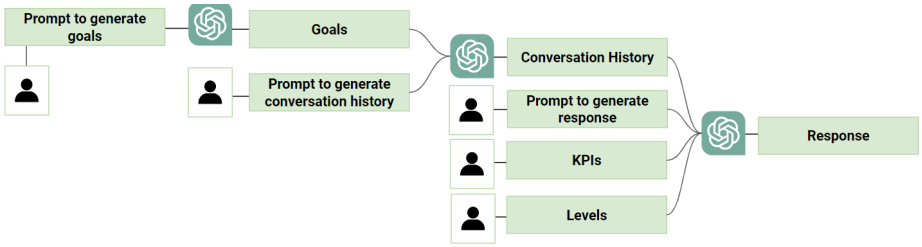


Fig. 1: Architecture of Dataset creation for experiment 1

the study's objectives. The structure, as depicted in figure 1, was designed to align each discussion with the KPIs. By generating the dataset in-house, a high level of control over content, consistency, and relevance was maintained throughout the dataset.

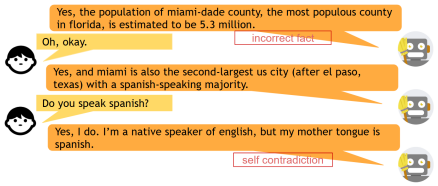


Fig. 2: Example of the response errors in a human-bot dialogue for Experiment 2. Adapted from Finch et al. [18].

In the second experiment, the ABC-Eval dataset from Finch et al. (2023) [18] was used as a benchmark for behavior error detection. This dataset includes 400 open-domain human-bot dialogues, reviewed by human annotators for specific dialogue errors. These annotations provided a robust benchmark for evaluating the system's performance across various conversational scenarios. These annotations were carried out by crowdworkers on the SurgeHQ platform, who received training using three curated conversations to ensure accurate identification of each dialogue behavior prior to participating in the annotation project. For example, in Figure 2, the three bot responses are annotated as 1, 0, 0 for the behavior "incorrect fact" (!Fac) and as 0, 0, 1 for the behavior "self-contradiction" (!Sel). This dataset provided a robust benchmark for evaluating the system's performance across various conversational scenarios.

GENERATE GOALS:

Generate an array of 6 conversation topics in a progression from simple to complex.

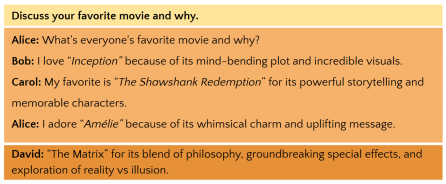
Example of a simple conversation topic:
Favorite Movies: "What's your favorite movie and why?"

Example of a more complex conversation topic:
Ethical Implications: What are the ethical implications of advancements like AI and biotechnology?

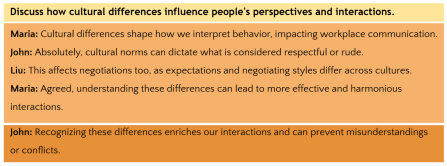
Fig. 4: Example prompt to generate the questions, which serve as goals for our discussion

3.1 Experiment 1

Goals We generated six discussion topics, ranging from simple to complex, to evaluate the dialogue system across diverse scenarios. Each topic was structured as a question, and dialogues were generated between four participants,



(a) **Information Seeking:** Participants respond directly to a question posed by Alice, sharing personal insights and preferences. The dialogue focuses on a straightforward exchange of information.



(b) **Exploratory:** Maria begins by discussing cultural differences and their impact on behavior interpretation. John and Liu expand on the topic, leading to a deeper exploration of how cultural norms affect workplace communication and negotiations

Fig. 3: Example of dialogues generated in different categories (Information-seeking and Exploratory)

categorized into *Information-Seeking* (IS) and *Exploratory* dialogues. Below are the list of generated goal:

1. Discuss your favourite movie and why?"
2. Discuss the most memorable place you've traveled to and what made it special?
3. Discuss the ethical implications of advancements like AI and biotechnology?
4. Discuss how cultural differences influence people's perspectives and interactions?
5. Discuss what hobbies or activities you enjoy in your free time?
6. Discuss your career goals and what steps are you taking to achieve them?

This task required significant effort and precision to ensure that the topics were both diverse and challenging, enabling us to thoroughly evaluate the performance and capabilities of the dialogue system across various scenarios. The dataset creation process followed the structure depicted in Figure 1.

Discussions After the *discussion goals* were established, each goal was used to generate dialogues between four participants. The discussions were predominantly categorized into *Information-Seeking* (IS) dialogues (see example in Figure 3a) and *Exploratory dialogues* (see example in Figure 3b). The former dialogues are structured primarily to acquire information from participants [49], while the later involve participants building upon the conversation initiated by a previous participant [28].

To generate these dialogues, prompts were designed to ensure that the conversations were relevant, coherent, and aligned with the evaluation criteria. See figure 13 in appendix for prompt to used to generate discussion.

Next Turn The last step was to generate a next turn in the conversation, which we will also call a *response turn* to all 6 discussions (see figure 5).Every discussion has to have a different response which should be evaluated in relation to the dialog, the given KPI

and a score level (see example in figures 7a and 7b below).

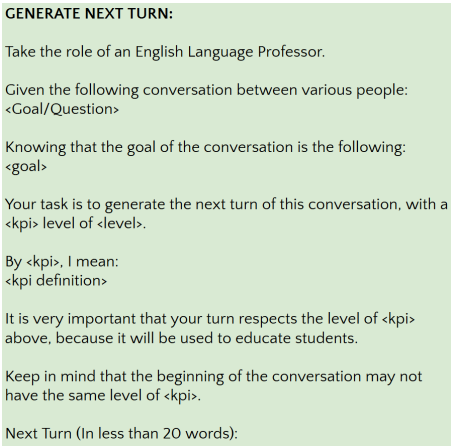


Fig. 5: Example of a prompt used to query GPT-4o to generate a response: Note: every word enclose

KPI Definitions The evaluation of dialogues was based on four key performance indicators (KPIs), selected to capture essential aspects of effective communication, emphasizing the importance of human-centric evaluation criteria in conversational AI [31][13]:

These KPIs were selected as they capture essential dimensions of dialogue quality that are challenging to measure but vital for assessing the performance of conversational AI systems. This is extensively discussed in works such as [53], [22] amongst others.

Score Levels: Each response was classified into a specific score level, ranging from 0 to 100%, in 20% increments. This approach allowed us to capture subtle differences in response quality, providing a more nuanced evaluation of the dialogue system's performance. The use of multiple score levels ensured that the evaluation could reflect variations in response quality more accurately, enabling a more detailed analysis of the system's strengths and weaknesses.

Model Settings: The GPT-4o model settings used throughout the experiments had the following settings: a temperature of 0.8, top-p set to 1.0, a maximum token usage of 40%, and no penalties or stop sequences. These settings were chosen to balance creativity and relevance, ensuring high-quality responses across various scenarios. These settings were carefully selected to balance creativity and relevance in the generated dialogues, ensuring that the model could produce good-quality responses across a range of scenarios.

Methodology To evaluate the generated dialogues, a survey was conducted with human participants. Four Google Forms were created, each focusing on one of the four KPIs. Each form initially contained six discussion goals, discussions, and corresponding responses. To enhance the reliability of human evaluations, a verification step was added, where responses at the same score level were duplicated. This allowed us to assess whether human evaluators could consistently identify responses that were rated similarly (with p_value of 0.347374, for $p = 0.05$ in the wilcoxon signed-ranked test). Following the human evaluation, GPT-4o was tasked with replicating the evaluation process using a carefully crafted prompt that included the discussion goals, dialogues, and individ-

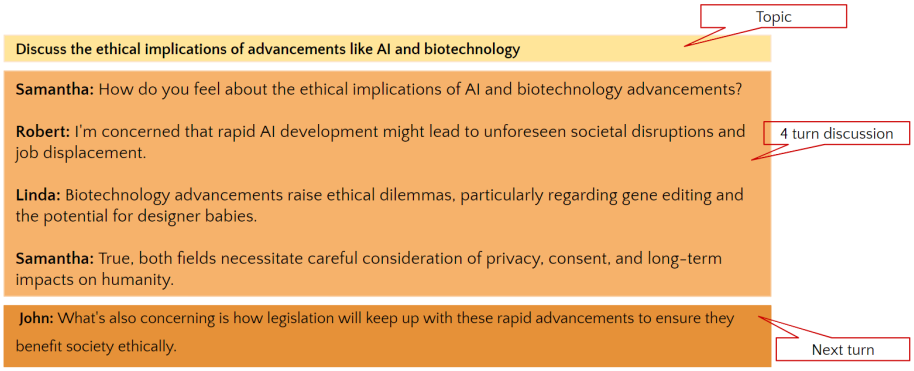
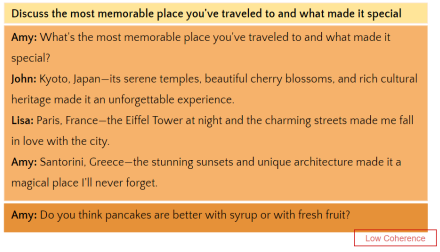
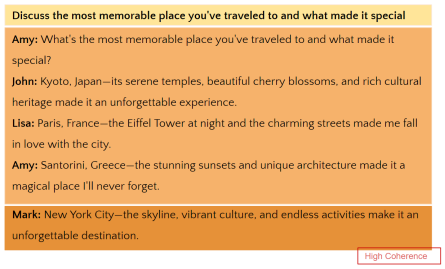


Fig. 6: Multi-party Dialogue Sample: Topic, 4-turn discussion and Next turn



(a) **Low Coherence:** The next turn introduces an unrelated topic, reducing coherence.



(b) **High Coherence:** All turns maintain relevant, cohesive responses.

Fig. 7: Example of next turns generated for a high- and low-level coherence.

ual responses. This setup allowed for a direct comparison between human and AI evaluations, providing insights into the alignment between the two.

Results

- **Human Responses:** The survey included 33 participants, with each participant providing unique responses across the four forms. This diversity in responses ensured a robust dataset for analysis, allowing us to evaluate the consistency and accuracy of human judgments.
- **GPT-4 Generated Responses:** For each human response, GPT-4o generated corresponding outputs. This parallel setup enabled a systematic comparison between human and AI responses, focusing on the consistency and alignment of ratings across all four KPIs. In total, 33 GPT-4 responses were generated, covering all four KPIs, which allowed for a comprehensive evaluation of GPT-4o's performance.

Terminology

Table 1: This table presents the Key Performance Indicators (KPIs) used for evaluating dialogue quality. It includes the definitions provided to both human participants and GPT models for assessment. Additionally, the "Aim" column specifies the purpose or objective of each KPI.

Name	Definition	Aim
Coherence	The degree to which the turn relates to the discussion.	To assess the relevance of a response to the ongoing discussion, ensuring that the dialogue remains focused and logical.
Innovation	The degree to which the turn adds new and creative ideas or solutions to the discussion, promoting novel approaches and perspectives.	To evaluate the introduction of new and creative ideas, which is critical for advancing the conversation and bringing fresh perspectives.
Concreteness	The degree to which the turn adds specific, tangible, and clear information or details to the discussion to enrich it.	To measure the degree to which a response provides specific, tangible details, enhancing the richness and clarity of the dialogue.
Goal Contribution	The degree to which the turn adds relevant and effective points to the discussion goal, advancing towards the overall objective or outcome.	To determine how effectively a response advances the overall objective of the discussion, contributing to achieving the intended outcome.

- **Human Rating**: Ratings assigned by human evaluators based on their opinion of the quality and relevance of each response.
- **GPT Rating**: Ratings assigned by GPT-4o based on its assessment of the quality and relevance of each response.
- **GPT Generation**: The predefined rating level used as a guideline for GPT4.o to generate the dataset responses.

Parameters Data analysis involved both linear regression and sigmoid fitting techniques to evaluate the correlation between human and GPT ratings. As shown in Figure 9, the linear model (green line) demonstrated a strong correlation coefficient of 0.997, indicating a significant linear relationship. However, the sigmoid function provided an even better fit, with a correlation coefficient of 1.00. The results highlight the binary nature of both human and GPT evaluations, where responses are often judged as either meeting or not meeting certain criteria, rather than on a continuous scale. This binary behavior aligns well with the sigmoid function, which effectively models step-like changes in the data.

KPI Analysis The analysis of the key performance indicators (KPIs)—Coherence, Innovation, Concreteness, and Goal Contribution—revealed strong correlations between human and GPT-4o ratings. These results can be seen in figure 9 and 8

- **Coherence**: There was a very high correlation between human and

Table 2: The table shows the correlation values between different ratings, demonstrating that GPT ratings closely align with human ratings across various KPIs, with particularly high correlations for Coherence and Goal Contribution.

KPI	Human vs. GPT Rating (Linear)	Human vs. GPT Rating (Sigmoid)	GPT Generation vs. Human Rating (Linear)	GPT Generation vs. Human Rating (Sigmoid)	GPT Generation vs. GPT Rating (Linear)	GPT Generation vs. GPT Rating (Sigmoid)
COHERENCE	0.99	1.0	0.92	1.0	0.90	1.0
INNOVATION	0.95	0.99	0.75	0.97	0.75	1.0
CONCRETENESS	0.87	0.96	0.47	0.72	0.71	0.96
GOAL CONTRIBUTION	0.99	1.0	0.74	0.96	0.67	0.95

Table 3: Summary of significance testing results comparing human ratings, GPT ratings, and GPT-generated responses.

HUMAN RATING VS GPT RATING		Normally Distributed (Shapiro-Wilk Test) >0.05	Paired t-test		Wilcoxon Signed-Rank Test	
No.	**KPI**	**Shapiro-Wilk Test Result**	**T-statistic**	**p-value (>0.05)**	**w-statistic**	**p-value (>0.05)**
1	**COHERENCE**	Yes – 0.56	-1.99	0.103		
2	**INNOVATION**	Yes – 0.83	-9.59	0.0002		
3	**CONCRETENESS**	Yes – 0.42	-1.19	0.29		
4	**GOAL CONTRIBUTION**	No – 0.0003			6.0	0.44

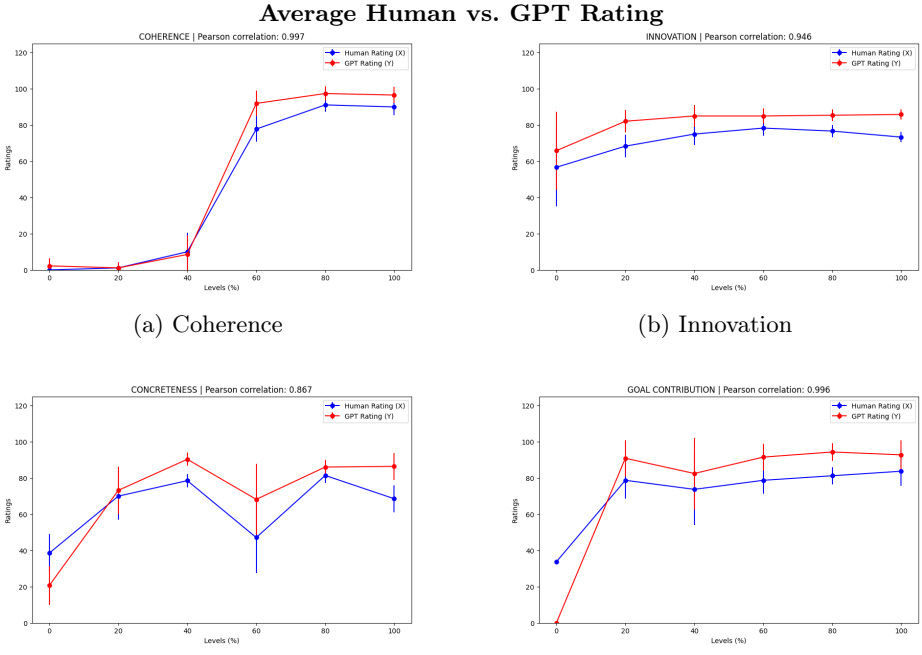
HUMAN RATING VS GPT GENERATION		Normally Distributed (Shapiro-Wilk Test) >0.05	Paired t-test		Wilcoxon Signed-Rank Test	
No.	**KPI**	**Shapiro-Wilk Test Result**	**T-statistic**	**p-value (>0.05)**	**w-statistic**	**p-value (>0.05)**
1	**COHERENCE**	Yes – 0.91	-0.67	0.53		
2	**INNOVATION**	Yes – 0.75	1.64	0.16		
3	**CONCRETENESS**	Yes – 0.38	1.04	0.34		
4	**GOAL CONTRIBUTION**	Yes – 0.92	1.99	0.10		

GPT RATING VS GPT GENERATION		Normally Distributed (Shapiro-Wilk Test) >0.05	Paired t-test		Wilcoxon Signed-Rank Test	
No.	**KPI**	**Shapiro-Wilk Test Result**	**T-statistic**	**p-value (>0.05)**	**w-statistic**	**p-value (>0.05)**
1	**COHERENCE**	Yes – 0.98	-0.04	0.97		
2	**INNOVATION**	Yes – 0.62	2.42	0.60		
3	**CONCRETENESS**	Yes – 0.49	1.93	0.11		
4	**GOAL CONTRIBUTION**	Yes – 0.80	2.13	0.09		

GPT-4o ratings, with a linear correlation of 0.99 and a perfect sigmoid correlation of 1.00. This indicates strong agreement in assessing the relevance of responses. This KPI is independent of the topic discussed and is straightforward to understand for both human and GPT.

– **Innovation**: For Innovation, the linear correlation between human and GPT-4o ratings was 0.95, with the sigmoid model improving to 0.99. The correlation between GPT-generated and human ratings was 0.75 (linear) and 1.00 (sigmoid), highlighting GPT-4o's strong alignment with human judgments in assessing creativity. This KPI might be influenced by the topic. Some topic being more appropriate to display innovation than others.

– **Concreteness**: Concreteness showed a linear correlation of 0.87 between human and GPT-4o rat-



(a) Coherence (b) Innovation

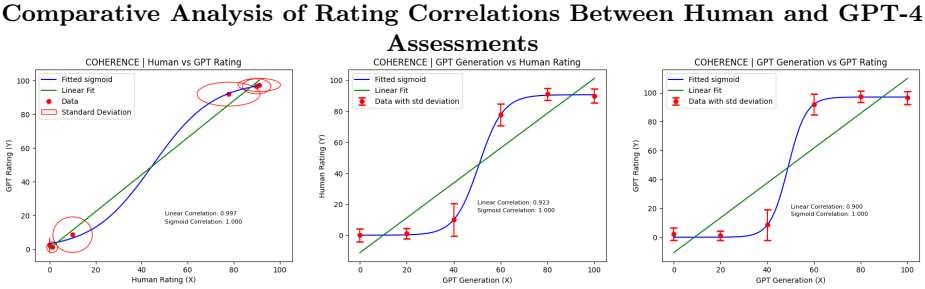
(c) Concreteness (d) Goal Contribution

Fig. 8: The chart displays the average ratings of humans and GPT across all four KPIs. Coherence has the highest correlation, although the response shows clearly a binary classification. Innovation seems to be very difficult to judge for both Human and LLM (the range of response is restricted to [50-90]). Concreteness has the lowest by correlation, between human and LLM nevertheless showing the same kind of response difficulty for both judges. Human and LLM shows also a binary behavior for goal contribution but with a bias towards high ratings.

ings, increasing to 0.96 with the sigmoid model. The correlation between GPT-generated and human ratings was 0.47 (linear) and 0.72 (sigmoid), indicating better alignment when non-linear aspects are considered. This KPI like the former one is prone to topic difference and might become very subjective in some cases, which explains the lower correlation.

– **Goal Contribution**: Goal Contribution had a linear correlation of 0.99 between human and GPT-4o ratings, with a perfect 1.00 in the sigmoid model. The correlation between GPT-generated and human ratings was 0.74 (linear) and 0.96 (sigmoid), reflecting strong agreement in evaluating how responses advance the discussion. Like the concreteness, this KPI is topic independent which is reflected in the high correlation.

Overall, these results show a high level of consistency between human and GPT-4o evaluations, particularly



(a) Human ratings vs. GPT-4 ratings (b) GPT-generated ratings vs. Human ratings (c) GPT-generated ratings vs. GPT-4 ratings

Fig. 9: The plots illustrate the linear and sigmoid correlations between different rating sources. (a) shows the correlation between Human ratings and GPT-4 ratings, (b) shows the correlation between GPT-generated ratings and Human ratings, and (c) shows the correlation between GPT-generated ratings and GPT-4 ratings. The fitted models and standard deviations are indicated by red circles.

when using a model that captures non-linear relationships.

Statistical Significance analysis To ensure the reliability of the results, statistical significance tests were performed. The Shapiro-Wilk test was used to determine whether the ratings followed a normal distribution. See table 3. Depending on the distribution results, either paired t-tests or Wilcoxon Signed-Rank tests were applied. For example, the Shapiro-Wilk test indicated a normal distribution for coherence, innovation, and concreteness, allowing for paired t-tests to compare human and GPT ratings. The results showed no significant differences in coherence and concreteness ratings, suggesting that GPT-4o's evaluations are consistent with human judgments. However, a significant difference was found in innovation ratings, indicating a potential area for improvement in GPT-4o's assessments. For the "Goal Contribution" KPI, where the data did not follow a normal distribution, the Wilcoxon Signed-Rank test showed no significant difference, further confirming the alignment between human and GPT ratings.

These statistical tests confirm that the observed similarities and differences between human ratings, GPT ratings, and GPT-generated responses are consistent and not due to random variation, enhancing the credibility of the results.

3.2 Experiment 2

Overview This experiment focused on assessing the quality of system-generated utterances by detecting errors, providing a clear metric for evaluating dialogue systems. Building on the work of Finch et al. [18], where GPT-3.5-turbo's error judgments were compared with human evaluations, we aimed to evaluate the error detection capabilities of the GPT-4.0-omni model using the same criteria.

Type of KPI	Description
Commonsense contradiction	The response misunderstands or contradicts common knowledge.
Incorrect fact	The response hallucinates or inaccurately presents encyclopedic or expert knowledge.
Redundant	The response inappropriately repeats information presented earlier in the dialogue.
Ignore	The response ignores what the user just said.
Irrelevant	The response interrupts the current topic of discussion by presenting unrelated information.
Partner contradiction	The bot contradicts or misremembers something the user said earlier in the dialogue.
Self Contradiction	The bot contradicts something it said earlier in the dialogue.

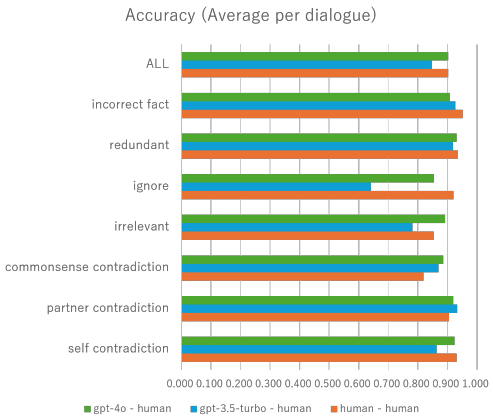
Fig. 10: List of KPIs.

Key Performance Indicators The accuracy of the responses was assessed based on specific error-related KPIs, as shown in Figure 10. Only the KPIs related to errors from the study by [18] were considered in this experiment.

Results The dataset from the previous study, containing 108 dialogue sets with evaluations by two human judges and GPT-3.5-turbo, was publicly available. We used the same prompts and generated new results with GPT-4.0-omni. Accuracy for each dialogue was calculated and averaged, with a significance test confirming the consistency of results down to three decimal places.

In cases of discrepancy between human judgments, the matching judgment with each GPT model was used for accuracy calculations. This sometimes resulted in higher accuracy for the GPT models compared to human-to-human comparisons. The results comparing GPT-4.0-omni, GPT-3.5-turbo, and human responses are summarized in the bar chart, highlighting the average accuracy and error rates for each KPI.

- **Overall Accuracy (ALL):** GPT-4.0-omni's accuracy was on par with human responses and significantly higher than GPT-3.5-turbo, showing its improved ability to identify errors.
- **Incorrect Facts:** Human evaluators were the most accurate in detecting incorrect facts, with significant differences observed between human and both GPT models, though the difference between GPT-3.5-turbo and GPT-4.0-omni was not significant. This was to be expected since fact checking generally difficult and knowledge dependent.
- **Redundant Responses:** Similar accuracy was observed across GPT-4.0-omni, GPT-3.5-turbo, and human responses in detecting redundancy, with no significant differences.
- **Ignoring User Input:** GPT-4.0-omni outperformed GPT-3.5-turbo but was still slightly less accurate than human evaluators.
- **Irrelevant Information:** GPT-4.0-omni surpassed both GPT-3.5-turbo and human evaluators in detecting irrelevant information.
- **Commonsense Contradictions:** GPT-4.0-omni had the highest accuracy for commonsense contradictions, though the difference with GPT-3.5-turbo was not significant.
- **Partner Contradictions:** GPT-3.5-turbo excelled in detecting partner contradictions, but the difference with GPT-4.0-omni was not statistically significant.
- **Self Contradictions:** Human responses were slightly more accurate in detecting self-contradictions, with GPT-4.0-omni results being nearly equivalent and not significantly different.



Statistical Significance (p-value) of Averaged Accuracy

	human vs GPT3.5	human vs GPT4o	GPT3.5 vs GPT4o
ALL	**0.00000**	0.98389	**0.00000**
incorrect fact	**0.03771**	**0.00062**	0.15471
redundant	0.13167	0.70332	0.21475
ignore	**0.00000**	**0.00000**	**0.00000**
irrelevant	**0.00001**	**0.00387**	**0.00000**
commonsense contradiction	**0.00260**	**0.00004**	0.28118
partner contradiction	**0.02553**	0.22959	0.27073
self contradiction	**0.00000**	0.52250	**0.00000**

Fig. 11: Accuracy result of experiment 2.

4 Conclusion

4.1 Experiment 1

Experiment 1 highlights the significant potential of GPT-based models in automating dialogue evaluation. The strong alignment between GPT ratings and human evaluations across KPIs such as Coherence, Innovation, Concreteness, and Goal Contribution, particularly when using a sigmoid model, suggests that GPT models can closely replicate human judgment. However, the differences observed between linear and sigmoid correlations, especially in Innovation and Concreteness, indicate areas where GPT models need refinement or where the topic of discussion might have an impact. These variations may stem from the subjective nature of these KPIs or inconsistencies in prompt design, pointing to the need for further research to address these challenges.

4.2 Experiment 2

Experiment 2 provides valuable insights into the comparative accuracy of GPT-4.0 and human evaluators in dialogue interactions. The analysis reveals that while GPT-4.0 excels in maintaining factual accuracy and commonsense reasoning, it still struggles with redundancy and self-contradiction. Conversely, humans are generally better at avoiding self-contradiction but are more prone to errors in commonsense and factual accuracy. These findings highlight areas for improvement in GPT models, emphasizing the need for continuous training and refinement to enhance the reliability and human-likeness of conversational agents.

5 Discussion

The results from both experiments show that GPT-4.0 has strong potential to replace human judgment in automated dialogue evaluation. Its strengths in maintaining factual accuracy and commonsense understanding make it a reliable tool for assessing conversations objectively. This marks a significant step forward in reducing reliance on human evaluators, particu-

larly for large-scale or repetitive evaluation tasks.

While humans excel at maintaining dialogue coherence and detecting nuanced inconsistencies, GPT-4.0 offers greater consistency in factual and logical reasoning. This suggests that AI is well-suited for the core tasks of dialogue evaluation. However, for highly detailed or subjective analyses, combining human expertise with AI's efficiency can provide an even more comprehensive assessment. The differences observed in response patterns—like the non-linear relationships noted in Experiment 1—highlight the complexity of human-like dialogue interactions. These findings underline the importance of improving evaluation methods to match this complexity, ensuring AI tools can meet the nuanced demands of conversational agents.

This study demonstrates that GPT-4.0 is a strong candidate for automating dialogue evaluation. By leveraging its strengths and refining its limitations, we can move closer to deploying AI as a standalone evaluator while reserving human input for specialized cases. This approach paves the way for more effective and scalable evaluation systems across diverse applications.

5.1 Challenges in Dialog Evaluation

Despite significant progress, challenges remain in evaluating dialogue systems. A major issue is the weak correlation between automated metrics and human judgments. Metrics like BLEU and ROUGE often do not align well with human perceptions of dialogue quality [20][43]. Additionally, evaluating multi-party dialogues presents unique challenges, as maintaining context and managing interactions among multiple participants add layers of complexity [20][51]. Future advancements in dialogue system evaluation will likely involve developing more inclusive metrics that integrate both human and automated evaluations. One promising approach is the use of prompting, which leverages large language models to generate and assess dialogue responses based on specific instructions or contexts [5][55]. This method allows for the simulation of various conversational scenarios and enables automated, scalable assessments that capture nuanced aspects of dialogue quality without requiring extensive annotated datasets [1][16].

5.2 Limitations and Future Works

Our study included a diverse range of human responses, from college students to university professors, which provided a broad spectrum of opinions on each KPI. However, focusing future research on experts in linguistics could lead to more consistent evaluations, as these experts might have more aligned interpretations of the KPIs. Incorporating multi-modal inputs, such as visual or auditory data, could also enhance the model's ability to generate more contextually appropriate responses by providing richer context.

Expanding the range of KPIs in future studies would represent a significant advancement in the field of NLP, particularly when working with complex models like GPT. This would allow for a more comprehensive evaluation of dialogue systems, ensuring that

they meet a broader set of criteria for human-like communication.

References

[1] Daniel Adiwardana et al. "Towards a human-like open-domain chatbot". In: *arXiv preprint arXiv:2001.09977* (2020).

[2] Libraria AI. *BERT and GPT Explained: How These Models Are Shaping the Future of AI*. Available at: `https://libraria.ai/bert-and-gpt-explained/`. 2023.

[3] Inshirah Ali AlMutairi and Ali Mustafa Qamar. "Evaluating Neural Dialogue Systems Using Deep Learning and Conversation History." In: *Journal of Artificial Intelligence (2579-0021)* 4.3 (2022).

[4] Emily M Bender et al. "On the dangers of stochastic parrots: Can language models be too big?¿¿". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 610–623.

[5] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[6] Tom B Brown et al. "Language models are few-shot learners. arXiv 2020". In: *arXiv preprint arXiv:2005.14165* 4 (2005).

[7] Paweł Budzianowski et al. "Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling". In: *arXiv preprint arXiv:1810.00278* (2018).

[8] *Call for Proposals - Dialog System Technology Challenge (DSTC) 12*. Association for Computational Linguistics. Available at: `https://www.aclweb.org`. 2024.

[9] Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. "MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 610–614.

[10] Ying Chen et al. "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety". In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. 2012, pp. 71–80. DOI: `10.1109/SocialCom-PASSAT.2012.55`.

[11] Mathias Creutz. "Open subtitles paraphrase corpus for six languages". In: *arXiv preprint arXiv:1809.06142* (2018).

[12] Bhavana Dalvi et al. "Explaining answers with entailment trees". In: *arXiv preprint arXiv:2104.08661* (2021).

[13] Jan Deriu et al. "Survey on evaluation methods for dialogue systems". In: *Artificial Intelligence Review* 54 (2021), pp. 755–810.

[14] David N Dobrin. "A new grammar checker". In: *Computers and the Humanities* 24 (1990), pp. 67–80.

[15] Michael Erdmann et al. "From manual to semi-automatic semantic annotation: About ontology-based text annotation tools". In: *Proceedings of the*

COLING-2000 workshop on semantic annotation and intelligent content. 2000, pp. 79–85.

[16] Yujie Feng et al. "Towards LLM-driven dialogue state tracking". In: *arXiv preprint arXiv:2310.14970* (2023).

[17] Sarah E Finch, James D Finch, and Jinho D Choi. "Don't Forget Your ABC's: Evaluating the State-of-the-Art in Chat-Oriented Dialogue Systems". In: *arXiv preprint arXiv:2212.09180* (2022).

[18] Sarah E Finch, Ellie S Paek, and Jinho D Choi. "Leveraging large language models for automated dialogue analysis". In: *arXiv preprint arXiv:2309.06490* (2023).

[19] Gartner. *58 Conversational AI Statistics You Can't Ignore*. Retrieved from Netomi. 2021. URL: `https://www.netomi.com/blog/conversational-ai-statistics`.

[20] Sarik Ghazarian et al. "Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7789–7796.

[21] Sukhpal Singh Gill et al. "Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots". In: *Internet of Things and Cyber-Physical Systems* 4 (2024), pp. 19–23.

[22] Herbert Paul Grice. "Logic and conversation". In: *Syntax and semantics* 3 (1975), pp. 43–58.

[23] Prakhar Gupta et al. "Investigating evaluation of open-domain dialogue systems with human generated multiple references". In: *arXiv preprint arXiv:1907.10568* (2019).

[24] Caroline Haist. "An Evaluation of Microsoft Word 97's Grammar Checker." In: (2000).

[25] Ryuichiro Higashinaka et al. "Integrated taxonomy of errors in chat-oriented dialogue systems". In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2021, pp. 89–98.

[26] Victor Hung et al. "Towards a method for evaluating naturalness in conversational dialog systems". In: *2009 IEEE international conference on systems, man and cybernetics*. IEEE. 2009, pp. 1236–1241.

[27] Frank Joublin, Antonello Ceravola, and Cristian Sandu. "Introducing Brain-like Concepts to Embodied Hand-crafted Dialog Management System". In: *arXiv preprint arXiv:2406.08996* (2024).

[28] Ben Kilby. "Dialogic Pedagogies: Defining and Analyzing Four Types of Dialogue in Education". In: *Analytic Teaching and Philosophical Praxis* 41.2 (2021), pp. 106–121.

[29] Simon Kloker, Matthew Bazanya, and Twaha Kateete. "I don't trust you (anymore)!–The effect of students' LLM use on Lecturer-Student-Trust in Higher Education". In: *arXiv preprint arXiv:2406.14871* (2024).

[30] Shivani Kumar et al. "Dialogue Agents 101: A Beginner's Guide to Critical Ingredients for De-

signing Effective Conversational Systems". In: *arXiv preprint arXiv:2307.07255* (2023).

[31] Jiwei Li, Will Monroe, and Dan Jurafsky. "Understanding neural networks through representation erasure". In: *arXiv preprint arXiv:1612.08220* (2016).

[32] Raymond Li et al. "Towards deep conversational recommendations". In: *Advances in neural information processing systems* 31 (2018).

[33] Ryan Lowe et al. "Towards an automatic turing test: Learning to evaluate dialogue responses". In: *arXiv preprint arXiv:1708.07149* (2017).

[34] Gale M Lucas et al. "Getting to know each other: The role of social dialogue in recovery from errors in social robots". In: *Proceedings of the 2018 acm/ieee international conference on human-robot interaction*. 2018, pp. 344–351.

[35] Khyati Mahajan, Sashank Santhanam, and Samira Shaikh. "Towards evaluation of multi-party dialogue systems". In: *Proceedings of the 15th International Conference on Natural Language Generation*. 2022, pp. 278–287.

[36] Michael McTear. *Conversational ai: Dialogue systems, conversational agents, and chatbots*. Springer Nature, 2022.

[37] AI Meta. "Introducing LLaMA: A foundational, 65-billion-parameter large language model". In: *Meta AI* (2023).

[38] Mikio Nakano et al. "A robot that can engage in both task-oriented and non-task-oriented dialogues". In: *2006 6th IEEE-RAS International Conference on Humanoid Robots*. IEEE. 2006, pp. 404–411.

[39] Jinjie Ni et al. "Recent advances in deep learning based dialogue systems: A systematic survey". In: *Artificial intelligence review* 56.4 (2023), pp. 3055–3155.

[40] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques". In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. 2002, pp. 79–86.

[41] Journal of Petroleum Technology. *As Hype Fades, LLMs Gaining Acceptance in Upstream as New-Age Research and Coding Tool*. Accessed: 2024-07-03. 2023. URL: `https://jpt.spe.org/as-hype-fades-llms-gaining-acceptance-in-upstream-as-new-age-research-and-coding-tool`.

[42] Sujan Reddy. "Automating Human Evaluation of Dialogue Systems". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. 2022, pp. 229–234.

[43] Mario Rodríguez-Cantelar et al. "Overview of robust and multilingual automatic evaluation metrics for open-domain dialogue systems at dstc 11 track 4". In: *arXiv preprint arXiv:2306.12794* (2023).

[44] Ananya B Sai et al. "Re-evaluating adem: A deeper look at scoring dialogue responses". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 6220–6227.

[45] De Witt T Starnes and Gertrude E Noyes. "The English Dictionary from Cawdrey to Johnson 1604-1755". In: (1991).

[46] K Sun et al. "DREAM: a challenge dataset and models for dialogue-based reading comprehension. CoRR abs/1902.00164 (2019)". In: *URL: http://arxiv.org/abs/1902.00164* (1902).

[47] Pirros Tsiakoulis et al. "Statistical methods for building robust spoken dialogue systems in an automobile". In: *Proceedings of the 4th applied human factors and ergonomics* (2012).

[48] Alex Vernon. "Computerized grammar checkers 2000: Capabilities, limitations, and pedagogical possibilities". In: *Computers and Composition* 17.3 (2000), pp. 329–349.

[49] Douglas Walton. "Types of dialogue and burdens of proof." In: *COMMA*. 2010, pp. 13–24.

[50] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection". In: *IEEE Access* 6 (2018), pp. 13825–13835. DOI: 10.1109/ACCESS.2018.2806394.

[51] Jimmy Wei et al. "Multi-party chat: Conversational agents in group settings with humans and models". In: *arXiv preprint arXiv:2304.13835* (2023).

[52] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1 (1966), pp. 36–45.

[53] George Yule and Gillian R Brown. *Discourse analysis*. Cambridge University Press, 1986.

[54] Saizheng Zhang et al. "Personalizing dialogue agents: I have a dog, do you have pets too?" In: *arXiv preprint arXiv:1801.07243* (2018).

[55] Yizhe Zhang et al. "Dialogpt: Large-scale generative pre-training for conversational response generation". In: *arXiv preprint arXiv:1911.00536* (2019).

Appendix

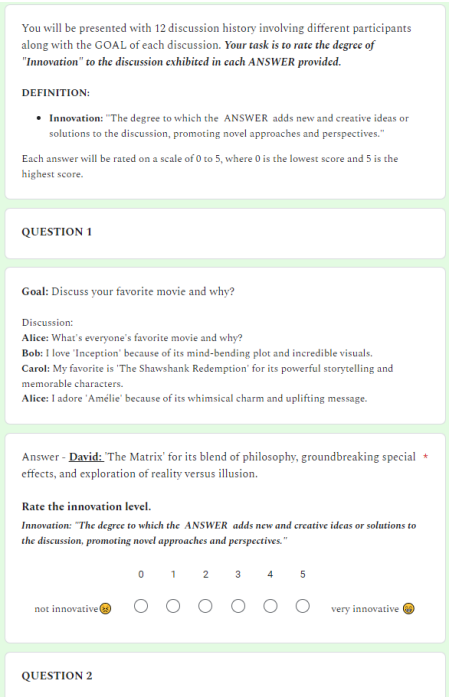


Fig. 12: Sample of the questionnaire for human evaluation discussed in the methodology section

```
GENERATE DISCUSSION:

Take the role of an English Language Professor.

Given this topic:
<goal>

Generate a 4 turn conversation with 3 participants (include names) - 1 sentence with a 20 word limit each

Example of output format should be:
[
 "Alice: What's everyone's favorite movie and why?",
 "Bob: I love "Inception" because of its mind-bending plot and incredible visuals.",
 "Carol: My favorite is "The Shawshank Redemption" for its powerful storytelling and memorable characters.",
 "Alice: I adore "Amélie" because of its whimsical charm and uplifting message."
]
```

Fig. 13: This is an example prompt for generating discussions, using the role of an English Language Professor to guide the conversation.