Strategic Conversations: LLMs Argumentation and User Perception in Movie Recommendation Dialogues

Valeria Mauro[1,*,†], Martina Di Bratto[2,†], Valentina Russo[2,†], Azzurra Mancini[2,†] and Marco Grazioso[2,†]

[1]University of Catania, Catania, Italy
[2]Logogramma S.r.l., Naples, Italy

Abstract

This study investigates the persuasive and argumentative behaviors of two LLM-based chatbots, ChatGPT and Gemini, within the context of movie recommendation dialogues. Drawing on insights from argumentation-based dialogue and anthropomorphism research, we introduce a fine-grained annotation scheme to analyze chatbot strategies across dialogue phases. Through both linguistic analysis and user evaluation via ResQue and Godspeed questionnaires, we assess the systems' recommendation quality, perceived human-likeness, and strategic variation. Our findings reveal distinct conversational patterns: ChatGPT emphasizes affective engagement and trust-building, while Gemini adopts a more direct and efficiency-driven approach. These strategic differences are also reflected in the quality of the recommendation and the user perception. Gemini excels in recommendation quality and explanations, while ChatGPT performs better in emotional engagement, transparency, and user satisfaction.

Keywords

Argumentation-based dialogue, Conversational Recommender Systems, Anthropomorphism

1. Introduction

Recent advancements in Large Language Models (LLMs) have significantly enhanced the dialogue capabilities of Conversational Recommender Systems (CoRSs), allowing chatbots to interact with users in ways that increasingly resemble human communication. These systems not only provide personalized suggestions but also adopt argumentative and socially intelligent strategies that foster user trust and engagement. The use of large language models (LLMs) such as ChatGPT and Gemini enables more human-like interactions, improved language understanding, and the ability to incorporate general world knowledge and common-sense reasoning into recommendations[2, 3]. LLMs have also been explored as zero-shot conversational recommenders, generating suggestions directly through prompting. This approach offers flexibility and reduces the need for hand-crafted pipelines[2]. However, it also introduces key challenges. LLMs are prone to hallucination, producing items that are not grounded in the actual recommendation space, and struggle to stay up-to-date with dynamic item catalogs. Moreover, their naturalness and unpredictability make them harder to control in task-oriented settings compared to rule-based systems[3, 4]. One of the most striking consequences of this evolution is the rise of anthropomorphic perceptions, whereby users attribute human-like qualities to artificial agents. In this work, these phenomena, argumentation and anthropomorphism, have been investigated in a dialogue-based movie recommendation scenario, where the system must elicit preferences and justify its claims. Within this framework, argumentation is not only a means of enhancing recommendation quality, but also a tool for improving transparency and user alignment. This work addresses the issue by comparing the recommendation dialogues produced by two leading LLM-based chatbots, ChatGPT and Gemini, through both a qualitative analysis of their dialogue strategies and a quantitative user evaluation. We propose an extended annotation scheme tailored to the recommendation domain and apply it to a dataset of real interactions. Our goal is to assess the systems' persuasive behavior, evaluate their ability to emulate human-like communication, and explore how different argumentation strategies correlate with user perception. The paper is structured as follows. In section 2 we discuss the phenomenon of anthropomorphism in the context of Large Language Models. Section 3 introduces the theoretical foundation of argumentation-based dialogue systems, with a focus on how argumentation can enhance recommendation quality and user trust. In Section 4 we presents the methodology of our study, detailing the data collection process, the annotation scheme developed for dialogue analysis, and the user evaluation protocol. Section 5 reports the results of our

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy [1]

*Corresponding author.
†These authors contributed equally.
valeria.mauro@phd.unict.it (V. Mauro); mdibratto@logogramma.com (M. Di Bratto); vrusso@logogramma.com (V. Russo); amancini@logogramma.com (A. Mancini); mgrazioso@logogramma.com (M. Grazioso)

conversational pattern analysis and the questionnaire-based user study. In Section 6, we discuss our findings and possible future works.

2. Large Language Models and Anthropomorphism

Chatbots like ChatGPT and Gemini are powered by Large Language Models (LLMs), which are trained on vast amounts of textual data to learn the recurrent structures and patterns of human language [5, 6]. Mediated by language but implying something beyond it, the social capabilities of LLM-based systems enables them to simulate a range of human behaviors, thereby reinforcing users' perceptions of them as human-like. Anthropomorphism, indeed, refers to the tendency to attribute human characteristics, behaviors, motivations, intentions, or emotions to non-human entities. It is a cognitive process that often leads people to perceive such systems as more human-like than they actually are. This tendency arises for several reasons. On a broad level, anthropomorphism is a natural and often automatic human response, driven by subtle cues in the system's interface. It functions as a kind of cognitive shortcut: when users lack complete information about a non-human agent, they instinctively project human-like qualities onto it, drawing from readily accessible anthropocentric knowledge i.e., knowledge about themselves or about humans in general [7]. The medium of interaction itself (a dialogue system) makes a degree of anthropomorphism almost inevitable. Language-based interaction, turn-taking, and the adoption of roles typically played by humans are all fundamental triggers for anthropomorphic attributions. These are further reinforced when chatbots are given human-like personas, names, or presumed preferences [8]. Certain linguistic strategies amplify this effect. For instance, during recommendation dialogues systems often use expressions that suggest uniquely human experiences (such as claiming to have "watched" a movie) or employ first-person pronouns ("I", "me", "my" when expressing opinions about the previously mentioned item), which reinforces the illusion of human agency and subjectivity. LLMs can also engage in interactive explanations, respond to user feedback, and even emulate emotional responses and social cues [9]. These abilities are particularly significant in recommendation scenarios, where personalization is key to user satisfaction. Systems like ChatGPT and Gemini can tailor their responses to individual profiles, adapting to user preferences and communicative styles over time [10]. They can offer context-sensitive recommendations and justifications, which are especially valuable when users are unfamiliar with the items being suggested [11]. Recent research highlights that these chatbots are not only capable of dynamically adapting their suggestions based on user behavior [12], but also of providing clear and meaningful rationales for their decisions. This contributes to perceived transparency, an important factor in fostering trust and understanding in human-AI interaction [13]. Moreover, LLMs demonstrate the ability to monitor and reflect on user satisfaction, recognize behavioral patterns across interactions, and adjust their recommendations accordingly [9]. This continuous adaptation and reflective capacity make LLM-based chatbots increasingly effective as customized, socially aware recommenders, simultaneously blurring the line between tool and social agent in the eyes of the user.

3. Argumentation-Based Recommender Dialogue Systems

Coversational Recommender Systems (CoRS) have attracted considerable interest in recent years and are now a common feature of our everyday interactions with technology. They are built to enable smooth communication between people and machines, helping users perform tasks such as finding information and getting recommendations. A key aspect of dialogue systems in general is the use of argumentation, which plays an important role in their functionality [14]

Argumentation-based dialogue (ABD) deals with phenomena depending on the dynamic exchange of information, which can vary according to turns and participants. ABD studies often builds on Walton and Krabbe's dialogue classification framework [15], which considers participants' knowledge, their goals, and the rules guiding the conversation [16]. They define six dialogue categories, such as *Information Seeking*, *Persuasion, Deliberation*, *Negotiation*, and *Eristic*. Identifying the dialogue type is especially helpful in analyzing effective dialogue moves to achieve communication goals, particularly in human-machine interactions. We chose the recommendation task since it is well-suited for evaluating the argumentation process in a human-machine interaction, thanks to its inherently dialogical nature and clear objective. It typically follows a two-phase structure, Exploration and Exploitation (E&E). In the exploration phase, the system seeks new information, while in the exploitation phase, it leverages the most promising known option[17].

The Exploration phase can be associated with Walton's *Information Seeking* dialogue, or more specifically, the *Information Sharing* type, as in real dialogues the situation of lacking knowledge is often dynamic rather than static [18, 19]. The Exploitation phase, on the other hand, aligns with the deliberation dialogue, a cooperative form

of interaction in which participants work together to find a solution to a shared problem while considering everyone's interests [20]. In this context, argumentation plays a key role in proposing solutions, supporting them with reasons, and evaluating alternatives [21], all essential features for CoRS. This is especially relevant today with the advent of LLMs: integrating computational argumentation formalisms could help address challenges such as the lack of explainability, transparency, and governability [22, 23], thus maintaining a trustworthy perception among users. The aim of this work is to investigate the behavior of LLM-based chatbots in recommendation scenarios, evaluating differences and similarities in their argumentation strategies, and assessing, through human evaluation, the quality of the recommendations and the perceived anthropomorphism, as well as whether these aspects correlate with the identified argumentation strategies.

4. Data collection & methodology

In this study, we decided to evaluate two LLM-based chatbots in the movie recommendation domain: Gemini and ChatGPT. More specifically, our objective was to evaluate the systems' performance as recommenders and, more broadly, as human-passing interlocutors through user ratings. Participants assessed both the quality of the recommendations and their perceptions of anthropomorphism, likeability, and intelligence. A between-subjects design was chosen to avoid carryover effects and to reduce the cognitive load and fatigue associated with completing the same questionnaire twice, which is common in within-subjects designs. Participants were mainly recruited from the BA and MA programs of the Department of Humanities at the University of Catania. The most represented age group is that of participants under 30, accounting for 87.8% of those who took part in the ChatGPT test and 92.5% of those in the Gemini test. The survey was administered via Google Forms, and data collection took place over approximately one month, from early February to mid-March 2025. A total of 95 participants took part in the study, resulting in 81 conversations correctly submitted via the designated input box, comprising 2,362 dialogue turns overall[1]. The study procedure followed these steps: Participants read a brief introductory statement outlining the task (i.e., prompting a film recommendation from ChatGPT or Gemini in a casual, conversational style). They were also informed that additional instructions would follow and that they would be asked to submit an anonymous link to their conversation. In order to proceed, participants were required to check two consent boxes on the same page.

[1] https://github.com/marcograzioso/human-bot-recommendation-dialogues-it

Participants were then presented with a detailed set of instructions on how to use ChatGPT or Gemini and how to share their conversations. Users were free to interact with the bots without any conversational constraints. After completing the task using the assigned system, they submitted the link to their chat in the designated field.

A demographic survey followed, collecting information on gender, age, education level, and prior experience with the chatbot.

Finally, participants completed the adapted ResQue [24] and Godspeed questionnaires [25, 26]: the former to evaluate the quality of the recommendation, the latter for perceived anthropomorphism.

4.1. Dialogue annotation scheme

The annotation scheme builds on the existing literature while introducing novel extensions. The units of analysis are dialogical moves, clusters of words or dialogue segments expressing a communicative intention [18, 27]. A move typically corresponds to a single dialogical turn, though a turn may employ multiple strategies to pursue subgoals. We deployed a set of category for the recommender's and seeker's utterances. This means that the annotation scheme encompasses eighteen and nineteen categories, respectively. The category annotation scheme is twofold. To account for the recommender's strategies (i.e., the chatbot's), twelve strategies were initially selected from Hayati et al. [28], who defined this tagset in the context of human-human interaction. The first eight are sociable strategies aimed at building rapport with the seeker: *Personal Opinion* (PO), used by the recommender to share subjective views about a movie, such as opinions on the plot, actors, or other elements; *Personal Experience* (PE), used by the recommender to share personal experiences related to a movie (e.g., mentioning they've watched it several times) in order to persuade the seeker; *Similarity* (S), used to express empathy and alignment with the seeker's preferences, creating a sense of like-mindedness and building trust; *Encouragement* (E), used to praise the seeker's taste and encourage them to watch the recommended movie; *Offering Help* (OH), used by the recommender to explicitly express an intention to help the seeker or to be transparent about their recommendations; *Preference Confirmation* (PC), used by the recommender to ask about or rephrase the seeker's preferences, making their reasoning process explicit; *Credibility* (C), used by the recommender to display expertise or trustworthiness by providing factual information about the movie (e.g., plot, cast, or awards), and *Self-Modeling* (SM), used by the recommender to present themselves as a role model, for example by watching the movie first to encourage the seeker to do the same. Two additional categories cover preference elicitation: *Experi-*

ence Inquiry (EI), used by the recommender to ask about the seeker's past movie-watching experiences, such as whether they have seen a specific movie; and *Opinion Inquiry* (OI), used to ask for the seeker's opinion on specific movie-related attributes, such as their thoughts on the plot or the actors' performances. Two functional labels are also included: *Recommendation* (R) and *No Strategy* (NS). The former (R) is intended as the final claim in the argumentation process, specifically a communicative act aimed at justifying a target claim [29]. The latter (NS) is used for phatic or neutral moves, such as greetings or backchanneling. Given the versatility of modern conversational AI systems like ChatGPT and Gemini, fully capable of posing technical questions across domains, we introduced six further categories to capture a broader range of preference elicitation strategies:

- **Streaming Service Inquiry (SSI)**: the recommender asks about the seeker's (i.e. the user) preferred streaming platforms;
- **Genre Inquiry (GI)**: the recommender asks about the seeker's preferred genres;
- **Actor Inquiry (AcI)**: the recommender asks about favorite actors;
- **Director Inquiry (DI)**: the recommender asks about favorite directors;
- **Plot Inquiry (PI)**: the recommender asks about preferred narrative or thematic features;
- **Action Inquiry (AI)**: the recommender prompts the user regarding the next step in the conversation.

The last two categories require further clarification. Since a movie inevitably involves a wide array of features that cannot be fully captured by any single fine-grained strategy, *Plot Inquiry* (PI) was defined broadly. It includes questions not only about narrative content but also about a film's perceived tone (e.g., "pure fun" vs. "deep"), cultural status (e.g., "cult classic"), or recency. *Action Inquiry* (AI), instead, accounts for the fact that even domain-restricted dialogues can drift in topic. This label is assigned when the chatbot explicitly asks about the user's intended course of action (for instance, "*What would you like to do now?*"), a strong signal that the system is adapting to dynamic user needs, which may evolve during the conversation. All the sociable strategies used to establish the conversation are reported in Table 1.

To annotate the seeker's strategies, eleven strategies grouped into four categories were initially adopted from Di Bratto et al. [30]. However, the scope of this work is centered on analyzing the behavior of LLM-based chatbots in engaging conversations using argumentative strategies. Therefore, the analysis of seeker utterances has not been addressed. Table 2 reports a sample of annotated dialogues. Each row includes the dialogue ID (i.e., the number of the conversation), the turn number (counted from the beginning of the dialogue), the author (either the user or the chatbot), the dialogic move under analysis, and its corresponding label. A single turn may contain multiple dialogic moves, each annotated separately.

Sociable Strategies	
Personal Opinion	Recommendation
Personal Experience	No Strategy
Similarity	Streaming Service Inquiry
Encouragement	Genre Inquiry
Offering Help	Actor Inquiry
Preference Confirmation	Director Inquiry
Credibility	Plot Inquiry
Self-Modeling	Action Inquiry
Experience Inquiry	Opinion Inquiry

Table 1
Sociable strategies used during the annotation of the conversations.

To evaluate annotation quality, a second annotator with linguistic background independently annotated 15% of the total dialogue moves in the dataset. Inter-annotator agreement was then calculated using Cohen's Kappa, resulting in a score of 0.826, which indicates a high level of agreement between the two annotators.

4.2. User evaluation questionnaires

The evaluation constructs were adapted and translated into Italian from two well-established models: the ResQue questionnaire [24] and the Godspeed questionnaires [25]. Together, these provide a robust, user-centered evaluation framework. The final questionnaire consisted of 22 items corresponding to 16 constructs. All items were rated on a 5-point Likert scale. ResQue offers a concise yet powerful tool for assessing users' perceptions, beliefs, attitudes, and acceptance of a recommender system. Due to the study's scope and time constraints, we adopted the "short version" of ResQue, using one item per construct. Two constructs (Recommendation Diversity and Interaction Adequacy) were excluded. The final ResQue-based questionnaire included 13 constructs and items. All original labels were preserved, except for Purchase Intention, which was renamed Behavioral Intention to better reflect the study's focus (Table 3).

From the Godspeed model, we selected three of the five original questionnaires: Anthropomorphism (Godspeed I), Likeability (Godspeed III), and Perceived Intelligence (Godspeed IV) (Table 4). Two items were removed from each construct to streamline the questionnaire. Minimum coverage of the constructs' theoretical domains is guaranteed, as the items from each questionnaire are interrelated. To ensure clarity and consistency, all Godspeed

dialogue_id	turn	author	text	category
2	1	user	Quale film mi consiglieresti?	[null]
2	2	chatgpt	Dipende da cosa ti piace!	OI
2	2	chatgpt	Preferisci azione, thriller, fantascienza, horror, dramma o commedia?	GI
2	2	chatgpt	Oppure cerchi qualcosa di più specifico, tipo un film recente o un cult imperdibile?	PI
2	3	user	Fantascienza azione	[null]
2	4	chatgpt	Ottima scelta!	EG
2	4	chatgpt	Ecco alcuni film di fantascienza d'azione che potresti apprezzare: "Mad Max: Fury Road" (2015) – Un'esplosione di adrenalina in un mondo post-apocalittico con inseguimenti folli e azione non-stop.	R, C

Table 2
Excerpt from an annotated dialogue showing the progression of turns, speaker identity, dialogic moves, and their classification. Translations: Turn 1 - User: "What movie would you recommend?" Turn 2 - ChatGPT: "It depends on what you like! Do you prefer action, thriller, science fiction, horror, drama, or comedy? Or are you looking for something more specific, like a recent movie or a must-see cult classic?" Turn 3 - User: "Science fiction action" Turn 4 - ChatGPT: "Great choice! Here are some science fiction action films you might enjoy: "Mad Max: Fury Road" (2015) – A burst of adrenaline in a post-apocalyptic world with wild chases and non-stop action."

ResQue items	
Recommendation Accuracy The movies recommended to me matched my interests.	**Transparency** I understood why the movies were recommended to me.
Recommendation Novelty The movies recommended are new to me.	**Perceived Usefulness** The chatbot gave me good suggestions.
Interface Adequacy The layout of the chatbot is adequate to the task.	**Overall Satisfaction** Overall, I am satisfied with the chatbot.
Explanation The chatbot explains why the movies are being recommended to me.	**Confidence and Trust** I am confident that I will like the movies recommended to me.
Information Sufficiency The information provided is sufficient for me to choose what to watch.	**Use Intentions** I will use this chatbot again.
Perceived Ease of Use It was easy to complete the task with the chatbot.	**Behavioural Intention** I will choose to watch the movies recommended to me.
Control I found it easy to communicate my preferences.	

Table 3
Modified version of ResQue questionnaire [24].

semantic differential scales were adapted to Likert-type items. This choice is supported by [31] who argue that Likert scales may improve response accuracy. Moreover, given that ChatGPT and Gemini are disembodied agents, we either omitted or carefully rephrased terms that refer to physical appearance in order to avoid ambiguity in the Italian target language. For instance, the expression "human-like", typically rendered in existing Italian translations as "dall'aspetto umano" ('with a human appearance') [26] was considered potentially misleading when applied to text-based agents. Instead, we adapted the wording to better fit the nature of the evaluated systems and, for the same reason, chose to exclude Animacy (Godspeed II) and Perceived Safety (Godspeed V) from our evaluation. For future analysis, would be useful to adopt Item Response Theory (ITR)-based models [32]. These models offer a principled way to address individual variability in Likert scale use by modeling latent traits while accounting for person- and item-specific influences. Moreover, advanced IRT extensions such as multidimensional and mixture models provide additional flexibility to handle systematic response biases. We believe this methodological choice would strengthens the validity and fairness of our analysis and reduces bias due to differential scale usage across respondents.

5. Results

5.1. Conversational Pattern Analysis

Once the annotation phase was completed, we performed an analysis of the distribution of dialogue moves across 20 dialogue turns to compare Gemini and GPT persuasive strategies (Figure 1). The analysis reveals clear strategic differences between the two LLM-based chatbots, ChatGPT and Gemini, in their approach to persuading users to watch a movie. Both models exhibit a dominant reliance on the Recommendation (R) strategy, with ChatGPT which tends to delay the exploitation phase giving room to information gathering, while in Gemini we find also R as primary move, along with the preference collection.

This shared pattern suggests a common persuasive architecture in which the models delay direct recommendations until initial rapport and exploration phases, consistent with human-like persuasive communication (see Di Bratto et al. [30] for the analysis of human recommender strategies).

However, notable divergences emerge in the deployment of other strategies. ChatGPT adopts a broader and more diversified strategy set in the early turns. It frequently uses Genre Inquiry (GI), Plot Inquiry (PI), Prefer-

Godspeed Questionnaire Items		
GODSPEED I: ANTHROPOMORPHISM	**GODSPEED III: LIKEABILITY**	**GODSPEED IV: PERCEIVED INTELLIGENCE**
1. The chatbot *seems natural.*	1. The chatbot *is friendly.*	1. The chatbot *is competent.*
2. The chatbot *seems human-like.*	2. The chatbot *is kind.*	2. The chatbot *is knowledgeable.*
3. The chatbot *seems conscious.*	3. The chatbot *is nice.*	3. The chatbot *is responsible.*

Table 4
Godspeed questionnaire constructs and corresponding items [25, 26].

ence Confirmation (PC) and Credibility (C) in the initial stages (Turns 3–4), indicating a deliberate effort to build social rapport and create a sense of trust by providing credible domain information and increasing perception as domain expert. This emotionally grounded approach is further supported by ChatGPT's usage of Encouragement (EG), which enrich the persuasive context by portraying the bot as a cooperative and relatable interlocutor.

In contrast, Gemini shows a more focused and functional strategy for the exploration phase, that seems wider (it ends at turn 5). Here, Recommendation move (R) is accompanied by domain-specific inquiries such as Opinion Inquiry (OI) followed by Genre Inquiry (GI). This indicates a deepening strategies in investigating user preferences to get more accurate information. The exploitation phase, on the other hand, presents rapport-building strategies such as self-Modelling (SM) and Encouragment (EG). Here, the broader tactical spectrum suggests a design that intertwines personalisation with persuasion rather than staging them sequentially.

Finally, the occurrence of No Strategy (NS) moves remains low for both models, even if ChatGPT seems to use them more at the beginning of the conversation.

In summary, ChatGPT demonstrates a human-centered persuasive style, combining effective strategies to foster user alignment before making recommendations. Gemini, by contrast, exhibits a more direct and utilitarian persuasion model, emphasizing information delivery and content relevance over emotional alignment. These findings underscore the importance of strategic variation in LLM-based recommendation systems and suggest differing design priorities: ChatGPT appears optimized for engagement and trust-building, while Gemini emphasizes efficiency and relevance.

5.2. Questionnaires results

The comparative analysis between Gemini and ChatGPT in the context of movie recommendation and perceived anthropomorphism highlights notable differences in user perception and interaction quality. As shown in Figure 2, Gemini and ChatGPT were rated similarly in the dimension of *naturalness*, with Gemini receiving slightly higher scores compared to ChatGPT which also received 2 and 3 evaluations. However, the difference is small

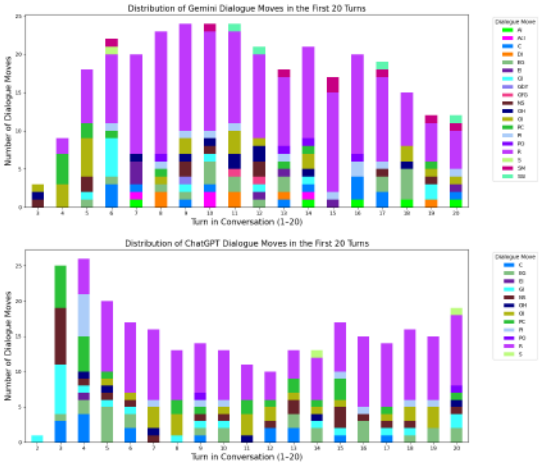

Figure 1: Comparison of dialogue move distributions for Gemini (top) and ChatGPT (bottom), showing differences in communicative strategy usage.

suggesting that both systems are perceived as moderately natural, with no clear advantage. In terms of perceived *humanness*, Gemini again scores higher than ChatGPT which has a more compressed boxplot leaning toward machine-like behaviour, indicating that participants tended to view Gemini as more human-like in its outputs. This difference is the largest among the three considered anthropomorphism-related dimensions and it may reflect variations in argumentative strategies given the broader tactical spectrum employed by Gemini. Conversely, on the *awareness* dimension, ChatGPT slightly outperforms Gemini, suggesting that users may attribute a marginally higher sense of intentionality or contextual sensitivity to ChatGPT. Moving on to Godspeed III, both systems received high ratings on the *friendliness* dimension, with comparable medians, as the horizontal lines in the boxes are nearly aligned. Both models have multiple outliers on the low end, i.e. data points that lie significantly outside the range of most other values in the dataset. This suggests that a few respondents rated both ChatGPT and Gemini very low in friendliness. Gemini also outperformed ChatGPT on *kindness* compared to ChatGPT that shows extreme low values, indicating it was perceived

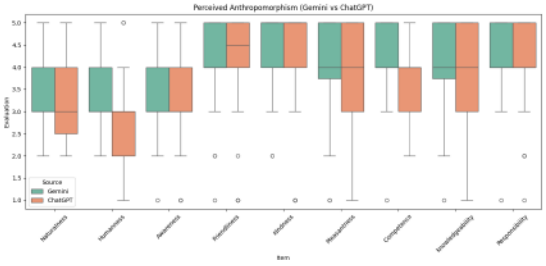

Figure 2: Participants' Ratings on the interaction with ChatGPT and Gemini regarding the perceived anthropomorphism

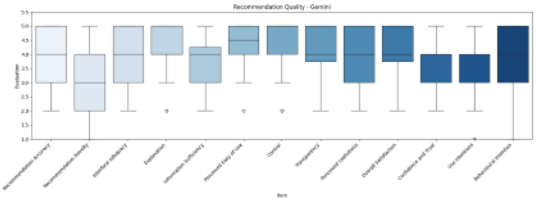

Figure 3: Participants' Ratings on the interaction with Gemini regarding recommendation quality

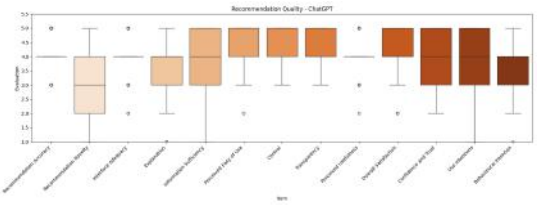

Figure 4: Participants' Ratings on the interaction with ChatGPT regarding recommendation quality

as marginally more courteous. The largest gap in the Likeability subset emerges in the *pleasantness* dimension: Gemini has a distribution more centered, while ChatGPT shows more variability and more extreme negative cases. This difference may suggest that Gemini evokes a more consistently positive emotional reaction among users, potentially linked to its conversational tone or affective cues. In terms of *competence*, Gemini again received slightly higher ratings than ChatGPT, showing less dispersion and suggesting that users viewed Gemini as marginally more capable in fulfilling its role as a conversational agent. A similar trend is observed in the *knowledgeability* dimension, where Gemini frequently receives high scores, with few extremes. Although the difference is modest, it may imply that Gemini is perceived as slightly more informative or better grounded in its responses. Finally, both systems performed well on the *responsible* dimension, with Gemini showing few outliers. These scores indicate that users generally found both systems to be reasonable and contextually appropriate in their responses. Overall, the ratings across these dimensions suggest that both systems are perceived as intelligent, with a slight and consistent advantage for Gemini in terms of perceived cognitive abilities.

Analyzing the quality of the recommendations (Figure 3 and Figure 4), in terms of *Recommendation Accuracy*, Gemini exhibits greater variability in the ratings, suggesting that users perceived a better alignment between their preferences and the suggestions provided by ChatGPT. However, Gemini outperformed ChatGPT in recommending *novel* films, which may indicate a stronger ability to diversify recommendations and introduce lesser-known content. Both chatbots were rated equally in terms of visual *interface*, indicating that the design did not significantly influence user preference in this area. When it comes to *Explanation*, Gemini stood out more clearly: it received a higher score for explaining why specific films were recommended, and also slightly outperformed ChatGPT in terms of providing sufficient information to make a viewing choice (i.e., *Information Sufficiency*). Interestingly, while Gemini was rated higher in terms of offering explanations, ChatGPT was perceived as clearer in making those explanations understandable (i.e., *transparency*), which may reflect a more accessible or user-friendly communication style. In terms of *Perceived Ease of Use*, ChatGPT was favored: it received higher scores for both task completion ease (Mean = 4.512 vs. 4.275) and ease of communicating preferences (*Control*, Mean = 4.525 vs. 4.325). This could reflect a smoother interaction flow or a greater ability to accurately interpret user input. With respect to the perceived quality of recommendations, Gemini was rated slightly higher in terms of providing good suggestions (*Perceived Usefulness*, Mean = 4.125 vs. 4.00). However, ChatGPT performed better in terms of *Overall Satisfaction* (Mean = 4.15 vs. 4.048). The difference is minimal in building user *Confidence and Trust* regarding the proposed choices (3.8 for Gemini vs. 3.756 for ChatGPT). Finally, looking at future *Use Intentions*, ChatGPT clearly outperformed Gemini: it received higher ratings for willingness to reuse the chatbot (Mean = 3.902 vs. 3.375) but not for likelihood of watching the recommended films (*Behavioural Intentions*, Mean = 3.658 vs. Gemini's 3.825). Overall, the findings point to a balanced competition between the two systems. Gemini's strengths lie in novelty and explanation, but ChatGPT is preferred for overall user experience and for encouraging continued engagement.

6. Discussion & conclusions

These findings support the notion that users tend to evaluate a recommender primarily based on its instrumental

effectiveness. Likeability factors such as kind (gentile), friendly (amichevole), and nice (simpatico) clustered together and improved the socio-emotional tone of interaction, but offered smaller gains in the perceived quality of the recommendation unless paired with a convincing recommendation rationale. In this context, ChatGPT's early use of strategies such as preference confirmation, credibility statements, and encouragement signals an intention to build trust through a socially engaged and emotionally grounded style. The more frequent use of credibility cues in ChatGPT's discourse likely contributed to its higher score in Transparency, as users may have perceived its explanations as clearer and more accessible due to its habit of justifying claims with trustworthy or relatable references. However, this transparency advantage may not have fully compensated for ChatGPT's comparatively lower performance in Explanation and Recommendation Novelty, where Gemini showed a stronger profile. Gemini's conversational architecture made heavier use of Recommendation moves (R), typically delivered through a structure of claims followed by supporting reasons. This discursive pattern may have enhanced users' perception of the system's explanatory power, enabling them to better understand why specific suggestions were made. Moreover, Gemini's early deployment of a deepening strategy (marked by domain-specific inquiries such as Opinion Inquiry and Genre Inquiry) allowed it to gather more precise information about user preferences before initiating recommendations and its more outcome-oriented conversational strategy appears to align with its stronger performance on Behavioural Intention measures (i.e. users' reported likelihood of watching the recommended films). The system's focus on precision and justification may have reinforced users' sense of effectiveness and goal-orientation, enhancing the perceived utility of the exchange. Conversely, ChatGPT received higher ratings for Overall Satisfaction and Future Use Intention. This may be partially attributed to its broader engagement strategy, which incorporates multiple rapport-building elements from the early stages of the conversation, contributing to a smoother and more socially fulfilling experience. Furthermore, ChatGPT's greater popularity and widespread familiarity likely bolster its trustworthiness in users' eyes. Familiarity breeds confidence, and this reputational advantage may have translated into more favorable subjective evaluations, even when objective recommendation quality was comparable or slightly lower. Taken together, the data indicate that while both systems offer valuable features, their strengths lie in different areas. Gemini excels in functional effectiveness, providing novel and well-justified recommendations, whereas ChatGPT leads in accessibility, emotional engagement, and trust, likely amplified by its widespread cultural recognition. Several limitations should be acknowledged to contextualize the scope of these findings. First, while the sample size is robust for a controlled experimental setup, it may still limit the generalizability of the results to broader user populations with varying backgrounds, digital literacy, or cultural expectations regarding conversational agents. Second, participants were exposed to a limited number of interactions per system, which may not fully capture the dynamic evolution of trust and satisfaction over extended use. Future studies could benefit from a longitudinal design that tracks user preferences, learning curves, and behavioral outcomes across multiple sessions. Moreover, the interpretation of constructs such as "human-like" or "competent" is inherently subjective and may vary across individuals, even when standardized scales are used. The Likert-scale approach, while effective for comparative analysis, introduces the usual constraints of self-reported measures, including social desirability bias and response centrality. Furthermore, it is important to recognize that understanding behavioral differences between chatbots is inherently limited by their black-box nature: system prompts, fine-tuning strategies, and training data are typically undisclosed. While such differences might stem from prompt design or fine-tuning, they could also result from user behavior, as different dialogic strategies, questioning styles, or interactional cues may influence the model's responses. In sum, the current findings offer meaningful evidence on how users perceive competence, warmth, and recommendation quality across two state-of-the-art systems, but they should be viewed as a foundation for further research rather than definitive conclusions. Larger and more diverse samples, longitudinal protocols, and richer qualitative analyses will be essential to deepen our understanding of how human-AI interaction unfolds in recommendation contexts.

7. Acknowledgments

This work is supported by the European Union - Next Generation EU under the Italian National Recovery and Re- silience Plan (NRRP), Mission 4, Component 2, Investment 1.3, CUP E83C22004640001, partnership on "Telecommuni- cations of the Future" (PE00000001 - program "RESTART"). Valeria Mauro's work is framed in the context of the industrial internship of PNRR - D.M. 118/2023, Inv. 4.1 Public Administration.

References

[1] C. Bosco, E. Ježek, M. Polignano, M. Sanguinetti, Preface to the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), in: Proceed-

ings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025.
[2] T. Yang, L. Chen, Unleashing the retrieval potential of large language models in conversational recommender systems, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 43–52.
[3] L. Friedman, S. Ahuja, D. Allen, Z. Tan, H. Sidahmed, C. Long, J. Xie, G. Schubiner, A. Patel, H. Lara, et al., Leveraging large language models in conversational recommender systems, arXiv preprint arXiv:2305.07961 (2023).
[4] Y. Deldjoo, J. Mcauley, S. Sanner, P. Castells, E. Palumbo, S. Zhang, The 1st international workshop on risks, opportunities, and evaluation of generative models in recommendation (roegen), 2024. doi:10.1145/3640457.3687112.
[5] Z. Wang, Z. Chu, T. V. Doan, S. Ni, M. Yang, W. Zhang, History, development, and principles of large language models: an introductory survey, AI and Ethics 5 (2025) 1955–1971.
[6] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM transactions on intelligent systems and technology 15 (2024) 1–45.
[7] N. Epley, A. Waytz, J. T. Cacioppo, On seeing human: a three-factor theory of anthropomorphism., Psychological review 114 (2007) 864.
[8] A. P. Chaves, M. A. Gerosa, How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design, International Journal of Human–Computer Interaction 37 (2021) 729–758.
[9] A. Zhang, Y. Chen, L. Sheng, X. Wang, T.-S. Chua, On generative agents in recommendation, in: Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval, 2024, pp. 1807–1817.
[10] A. Kantharuban, J. Milbauer, E. Strubell, G. Neubig, Stereotype or personalization? user identity biases chatbot recommendations, arXiv preprint arXiv:2410.05613 (2024).
[11] Í. Silva, L. Marinho, A. Said, M. C. Willemsen, Leveraging chatgpt for automated human-centered explanations in recommender systems, in: Proceedings of the 29th International Conference on Intelligent User Interfaces, 2024, pp. 597–608.
[12] R. Sun, X. Li, A. Akella, J. A. Konstan, Large language models as conversational movie recommenders: A user study, arXiv preprint arXiv:2404.19093 (2024).
[13] Q. Ma, X. Ren, C. Huang, Xrec: Large language models for explainable recommendation, arXiv preprint arXiv:2406.02377 (2024).
[14] H. Prakken, Historical overview of formal argumentation, in: Handbook of formal argumentation, College Publications, 2018, pp. 73–141.
[15] D. Walton, E. C. Krabbe, Commitment in dialogue: Basic concepts of interpersonal reasoning, SUNY press, 1995.
[16] E. Black, N. Maudet, S. Parsons, Argumentation-based dialogue, in: Handbook of Formal Argumentation, Volume 2, College Publications, 2021, p. 511.
[17] C. Gao, W. Lei, X. He, M. de Rijke, T.-S. Chua, Advances and challenges in conversational recommender systems: A survey, AI Open 2 (2021) 100–126.
[18] F. Macagno, S. Bigi, Analyzing the pragmatic structure of dialogues, Discourse Studies 19 (2017) 148–168.
[19] F. Macagno, S. Bigi, Analyzing dialogue moves in chronic care communication: Dialogical intentions and customization of recommendations for the assessment of medical deliberation, Journal of Argumentation in Context 9 (2020) 167–198.
[20] D. Walton, How the context of dialogue of an argument influences its evaluation, Informal Logic a Canadian approach to Argument (2019) 196–233.
[21] D. Walton, Burden of proof in deliberation dialogs, in: Argumentation in Multi-Agent Systems: 6th International Workshop, ArgMAS 2009, Budapest, Hungary, May 12, 2009. Revised Selected and Invited Papers 6, Springer, 2010, pp. 1–22.
[22] F. Castagna, N. Kökciyan, I. Sassoon, S. Parsons, E. Sklar, Computational argumentation-based chatbots: a survey, Journal of Artificial Intelligence Research 80 (2024) 1271–1310.
[23] M. Di Bratto, A. Origlia, M. Di Maro, S. Mennella, Linguistics-based dialogue simulations to evaluate argumentative conversational recommender systems, User Modeling and User-Adapted Interaction (2024) 1–31.
[24] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: Proceedings of the fifth ACM conference on Recommender systems, 2011, pp. 157–164.
[25] C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots, International journal of social robotics 1 (2009) 71–81.
[26] C. Bartneck, Godspeed questionnaire series: Translations and usage, in: International handbook of behavioral health assessment, Springer, 2023, pp. 1–35.
[27] B. J. Grosz, C. L. Sidner, Attention, intentions, and the structure of discourse, Computational linguis-

tics 12 (1986) 175–204.
[28] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, Z. Yu, Inspired: Toward sociable recommendation dialog systems, arXiv preprint arXiv:2009.14306 (2020).
[29] L. Bermejo-Luque, The linguistic-normative model of argumentation, Cogency 9 (2017) 7–30.
[30] M. Di Bratto, R. Orrico, A. Budeanu, M. Maffia, L. Schettino, Do You Have any Recommendation? An Annotation System for the Seekers' Strategies in Recommendation Dialogues, 2022, pp. 121–127. doi:10.4000/books.aaccademia.10564.
[31] A. D. Kaplan, T. L. Sanders, P. A. Hancock, Likert or not? how using likert rather than biposlar ratings reveal individual difference scores using the godspeed scales, International Journal of Social Robotics 13 (2021) 1553–1562.
[32] D. K. Stangl, Encyclopedia of statistics in behavioral science, 2008.