- UNIVERSITY OF CALIFORNIA

Santa Barbara

Effects of Visual Representation and Recommendation Bias in Conversational Recommender System

A Thesis submitted in partial satisfaction of the
requirements for the degree Master of Science
in Electrical and Computer Engineering

by

Cheng-Yan Wu

Committee in charge:

Professor Tobias Höllerer, Co-Chair

Professor Luke Theogarajan, Co-Chair

Professor Xifeng Yan

Professor Michael Beyeler

December 2023

The thesis of Cheng-Yan Wu is approved.

Michael Beyeler

Xifeng Yan

Luke Theogarajan, Committee Co-Chair

Tobias Höllerer, Committee Co-Chair

December 2023

ABSTRACT

Effects of Visual Representation and Recommendation Bias in Conversational Recommender System

by

Cheng-Yan Wu

This study explores the integration of Embodied Conversational Agents (ECAs) with Conversational Recommender Systems (CRS), focusing on the impacts of visual representation and recommendation bias. Leveraging an open-source Large Language Model alongside Nvidia Audio2Face and Unreal Engine 5, this study developed ECAs to assess their influence on user interaction in CRS. A 2x2 between-subjects study with 53 participants examined interactive avatars versus animated icons, evaluating their roles in enhancing recommendation persuasiveness and shaping user decision-making. Results indicate that while interactive avatars significantly boost user engagement, their influence on recommendation persuasiveness is minimal. Conversely, the presence of recommendation bias within CRS significantly impacts user opinions, highlighting its crucial role in CRS design. These findings underscore the importance of balancing visual innovation with ethical recommendation strategies in CRS, offering insights for advancing user-centered system development and informing future research.

TABLE OF CONTENTS

I. Introduction

Traditional recommender systems, primarily based on historical user data, often face limitations in scenarios involving high-involvement products, such as automobiles, where users typically lack a history of prior purchases to infer interests [8]. In response to these limitations, the advent of Large Language Models (LLMs) such as ChatGPT[1] and Gemini [2]marks a pivotal shift in Conversational Recommender Systems (CRS). These models enable multi-turn dialogues that more accurately capture user interests in real-time, even in the absence of extensive historical data. This advancement transforms the recommendation process into a dynamic and personalized interaction, representing a significant development beyond the capabilities of traditional recommender systems.

Despite the advancements in LLM-based CRS, the integration of visual representation, ranging from interactive avatars to animated icons, and its consequent impact on user behavior, especially in terms of engagement and persuasiveness, is an area that remains underexplored. This study aims to fill this gap by investigating the effects of such visual enhancements on CRS. We hypothesize that these visual elements can significantly influence user interaction, making the recommendation process more engaging and potentially more persuasive.

This study utilizes open-source Large Language Models (LLMs), focusing on privacy and customization. Open-source LLMs offer transparency and local execution capabilities, enhancing privacy by keeping data within the user's environment and allowing visibility into the model's workings. This approach aligns with the increasing emphasis on

[1] https://openai.com/blog/chatgpt

[2] gemini_1_report.pdf (storage.googleapis.com)

user trust and regulatory compliance in technology. Moreover, open-source models provide adaptability, enabling tailored functionality to meet specific user needs and preferences.

This study has the following objectives:

1. To leverage open-source LLM, providing a privacy-conscious alternative to proprietary models while maintaining high-quality user interactions.
2. To develop an interactive avatar framework that produces naturalistic facial expressions and vocal responses, avoiding the uncanny valley effect [20] during user-chatbot interactions.
3. To design and conduct a user study to evaluate the effects of combining open-source LLMs with interactive avatar technologies on user engagement and the persuasiveness of recommendations in CRS.

The contribution of this research is multifaceted, aiming to advance CRS design towards greater effectiveness, ethical considerations, and user-centeredness. By exploring the interplay between open-source LLMs, interactive avatars, and user engagement and persuasiveness, this study seeks to provide valuable insights and practical guidelines for enhancing the CRS experience.

II. Related Work

A. Conversational Recommender Systems

Traditional recommender systems, relying heavily on user behavior history, often encounter limitations in sparse data scenarios and lack user guidance clarity [8]. The advent of LLMs in CRS has revolutionized this domain by enabling more nuanced user-system interactions through dynamic, multi-turn dialogues [7]. These advancements effectively address the cold-start problem in recommendations and enhance the overall user experience.

Recent studies have focused on optimizing CRS architectures to integrate conversational and recommendation modules more effectively. Li et al. emphasized long short-term planning in CRS, aligning user preferences with recommendations more accurately [14]. Additionally, the role of LLMs in enhancing user interaction and personalization in CRS cannot be overstated. LLMs facilitate the generation of contextually relevant recommendations, tailoring the experience to individual users [7]. Furthermore, the evolution of CRS necessitates a reevaluation of existing assessment methodologies. Wang et al. have highlighted the need for rethinking evaluation techniques in the era of LLM-integrated CRS, focusing on interaction quality, recommendation accuracy, and user trust [25].

The trend towards interactive and explainable CRS using LLMs is also gaining significant traction in the field. These systems aim to enhance user trust and system transparency by providing clear and understandable recommendations. This aspect of CRS development is crucial for user acceptance and the effectiveness of the system [4]. The integration of LLMs in CRS has, therefore, not only enhanced the capabilities of these systems but also opened new avenues for research and development in creating more user-centric, trustworthy, and transparent recommendation systems.

B. Embodied Conversational Agents

Embodied Conversational Agents (ECAs) are integral to human-computer interaction, blending visual representation with conversational abilities to elevate user engagement. These agents, typically depicted as realistic avatars, facilitate human-like interactions, finding increasing use in sectors like customer service and digital health [3, 22, 24]. Their ability to produce naturalistic facial animations is critical, as these animations significantly contribute to immersive experiences, strengthening the connection between users and agents. Resources like the BEAT dataset are crucial for developing these animations, allowing for detailed exploration of human-avatar interactions [15].

In addition, the integration of AI and advanced computer graphics in creating Digital Human Avatars for Interactive Virtual Co-presence environments highlights the potential of ECAs to enhance the visual quality of conversational agents, offering lifelike and expressive avatars [12]. Zhu et al.'s study underscores this, showing how avatar representation and mixed reality settings influence user interactions and experience. Their findings reveal that virtual reality environments prompt more spatial references in conversations with avatars, indicating a user's perception of being in the avatar's space. Moreover, interactions with human avatars improved users' recall, suggesting a significant impact of human-like avatars on memory and experience in mixed reality [28].

This study explores the integration of Embodied Conversational Agents (ECAs) with Large Language Model-driven Conversational Recommender Systems (CRS), examining how ECAs' visual and interactive features influence the persuasiveness of CRS in real-world decision-making scenarios. This approach seeks to enhance the effectiveness of CRS by leveraging the advancements in human-avatar interactions.

III. System Design

The system design section outlines the architecture developed to enhance the Conversational Recommender System (CRS) and Embodied Conversational Agent (ECA) within this study. By integrating OpenAI's Whisper for Speech Recognition, Synthia70B-v1.2b as the Large Language Model, VITS for Text-to-Speech, Omniverse Audio2Face for audio-driven animation, and Unreal Engine 5 for rendering, we've constructed a robust framework. These components collectively enable a seamless and dynamic interaction in CRS by processing and synthesizing speech, text, and visuals in real-time.

A. Speech Recognition: OpenAI Whisper

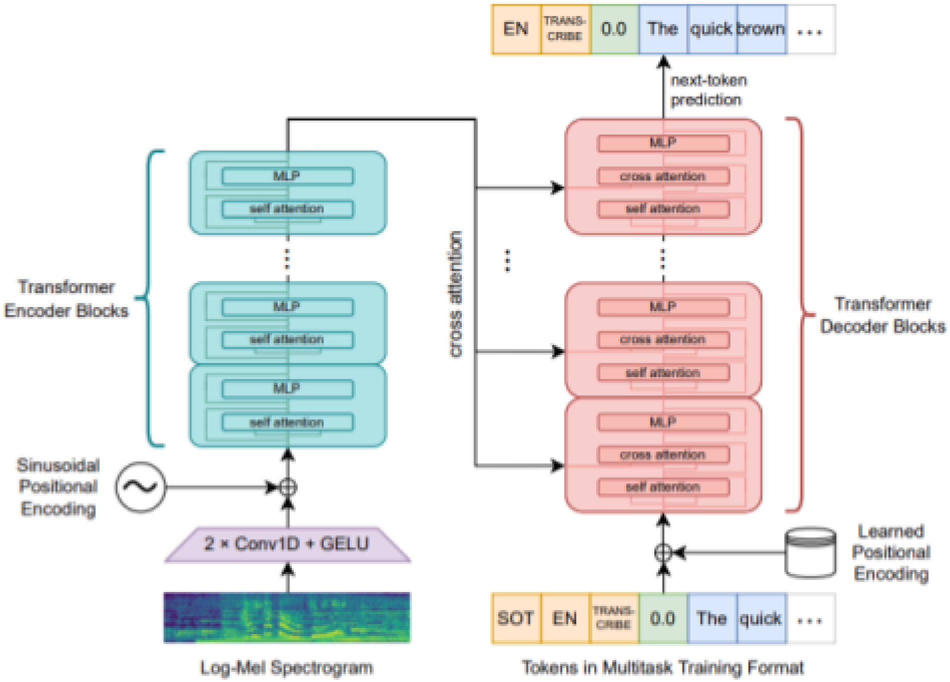


Figure 1: OpenAI Whisper's Encoder-Decoder Architecture [1]: Implements a sequence-to-sequence Transformer model, converting input audio into spectrogram form, processed by the encoder and decoded for text prediction.

Whisper [19], developed by OpenAI, is a Transformer-based sequence-to-sequence model, meticulously trained on 680,000 hours of supervised, multilingual, and diverse acoustic data from the web. This comprehensive training enhances its robustness and accuracy in recognizing speech across different languages, accents, and in the presence of background noise. In its inference process, the encoder analyzes incoming audio data, and the decoder, informed by the context of previously transcribed text, predicts subsequent words. This method ensures a more accurate transcription. For our research, we have selected the smaller, English-only model with 244M parameters, optimized for our speech recognition tasks.

B. Large Language Model: Synthia70B-v1.2b

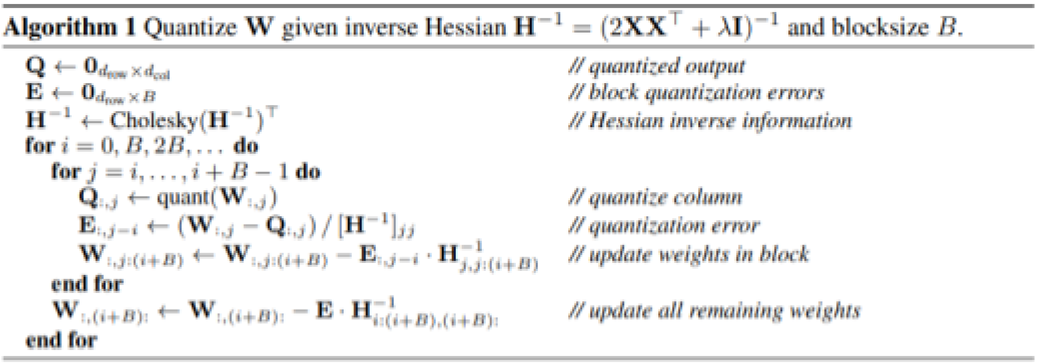
Synthia70B-v1.2b[3] leverages the foundational structure of Llama2-70B [22] , enhancing its instruction following and long-form conversation capabilities through fine-tuning on Orca-Style datasets [16]. This derivative model significantly advances in handling interim and concise conversation responses, as well as role-playing scenarios, building upon the robust base provided by Llama2-70B.

Llama 2 is a comprehensive collection of pretrained and fine-tuned large language models (LLMs) optimized for dialogue use cases, encompassing Llama 2-Chat models with parameters scaling from 7 billion to 70 billion. It employs an auto-regressive language model built upon an optimized transformer architecture. Specifically, Llama 2 uses the standard transformer architecture [1], applies pre-normalization using RMSNorm [26], utilizes the SwiGLU activation function [18], and incorporates rotary positional embeddings [9]. The training regimen consists of pretraining and fine-tuning stages, utilizing an auto-regressive transformer with data cleaning, updated data mixes, and increased context length during pre-training. This process saw a 40% increase in total tokens used for training, all performed on A100-80GB type hardware with carbon emissions offset by Meta's sustainability program. Llama 2 has a sequence length of 4096 tokens, allowing it to process up to around 1000 words, aligning well with our experiment design as the conversation usually wouldn't exceed this limit. The fine-tuning stage, crucial for aligning Llama 2-Chat with human preferences for helpfulness and safety, implements supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) alongside techniques like rejection sampling and Proximal Policy Optimization (PPO). Training data comprises 2

[3] README.md · migtissera/Synthia-70B-v1.2b at main (huggingface.co)

trillion tokens from publicly available sources for pretraining and publicly available instruction datasets plus over one million new human-annotated examples for fine-tuning, with no Meta user data included. Llama 2 emphasizes responsible LLM development, offering a responsible use guide and code examples to ensure safe deployment. This model represents a significant stride in language model advancement, showcasing pretrained and fine-tuned LLMs with enhanced performance in dialogue use cases, with a detailed exposition on the architecture, training data, and methodologies employed, accentuating the responsible development and safety enhancements integral to the models.

1. GPTQ Quantization



Algorithm 1 Quantize $\mathbf{W}$ given inverse Hessian $\mathbf{H}^{-1} = (2\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}$ and blocksize B.

```
$\mathbf{Q} \leftarrow \mathbf{0}_{d_{\text{row}} \times d_{\text{col}}}$                    // quantized output
$\mathbf{E} \leftarrow \mathbf{0}_{d_{\text{row}} \times B}$                         // block quantization errors
$\mathbf{H}^{-1} \leftarrow \text{Cholesky}(\mathbf{H}^{-1})^\top$                   // Hessian inverse information
for $i = 0, B, 2B, \ldots$ do
    for $j = i, \ldots, i + B - 1$ do
        $\mathbf{Q}_{:,j} \leftarrow \text{quant}(\mathbf{W}_{:,j})$                // quantize column
        $\mathbf{E}_{:,j-i} \leftarrow (\mathbf{W}_{:,j} - \mathbf{Q}_{:,j}) \,/\, [\mathbf{H}^{-1}]_{jj}$      // quantization error
        $\mathbf{W}_{:,j:(i+B)} \leftarrow \mathbf{W}_{:,j:(i+B)} - \mathbf{E}_{:,j-i} \cdot \mathbf{H}^{-1}_{j,j:(i+B)}$   // update weights in block
    end for
    $\mathbf{W}_{:,(i+B):} \leftarrow \mathbf{W}_{:,(i+B):} - \mathbf{E} \cdot \mathbf{H}^{-1}_{i:(i+B),(i+B):}$   // update all remaining weights
end for
```

Fig 2. GPTQ Algorithm Pseudocode

Utilizing GPTQ quantization [6], the model size of Synthia70B-v1.2b was effectively compressed from 16-bits to 4-bits, ensuring minimal loss in accuracy as depicted in Fig 4 on the LAMBDA benchmark for zero-shot task performance. The memory footprint was reduced from 140GB to 45GB without significant performance degradation. The pseudocode for GPTQ, provided in Fig 3, demonstrates the iterative process of quantizing columns of the model, computing quantization errors, and updating weights in blocks leveraging the inverse Hessian information. Although Fig 4 primarily showcases the performance of OPT and BLOOM models, the observed trends are expected to similarly manifest in Llama-based models, substantiating the efficacy of GPTQ quantization in this context.

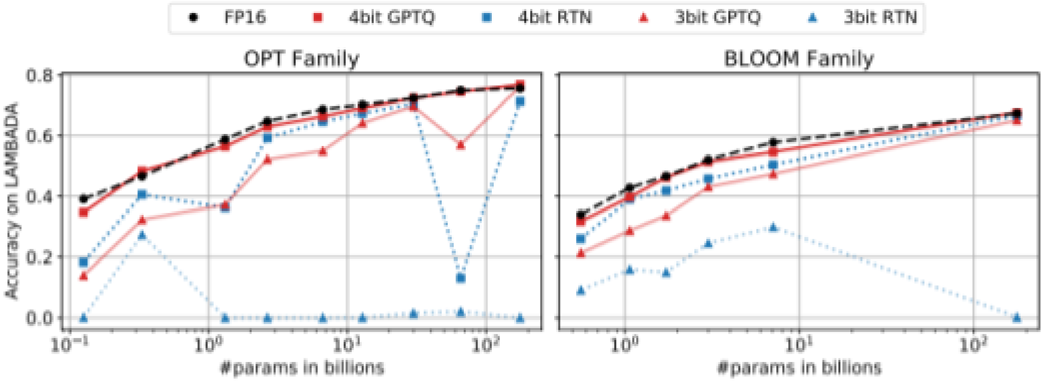


Fig 3. GPTQ Accuracy

2. ExLlamaV2: Fast and Memory-Efficient Inference

ExLlamaV2[4] is engineered to accelerate and optimize the inference process for large language models on contemporary consumer GPUs. The architecture incorporates a new quantization format, "EXL2," which permits a blend of quantization levels within a model, balancing between 2 and 8 bits per weight. This feature contributes significantly to

[4] GitHub - turboderp/exllamav2: A fast inference library for running LLMs locally on modern consumer-class GPUs

enhancing speed and reducing memory consumption. Quantization is pivotal, as it can diminish memory usage by 2x or 3x with 8-bit or 4-bit quantization, respectively.

Synthia-70b-v1.2b 4-bits GPTQ, ExLlamaV2 demonstrated a 2.5x speed up compared to Hugging Face (HF). The comparative performance and the ability to handle large models efficiently underscore ExLlamaV2's promise in significantly advancing the field of inference optimization.

C. Text-to-Speech: VITS

The system transforms the generated response text into spoken words using a text-to-speech (TTS) module. VITS [11] was selected for its ability to deliver real-time performance and produce a more natural voice output. Unlike other models that require distinct components for generating spectrograms and vocoding, VITS offers an integrated solution. It employs a conditional variational autoencoder coupled with adversarial learning, streamlining the synthesis process. VITS utilizes a conditional variational autoencoder architecture, comprising a posterior encoder, prior encoder,

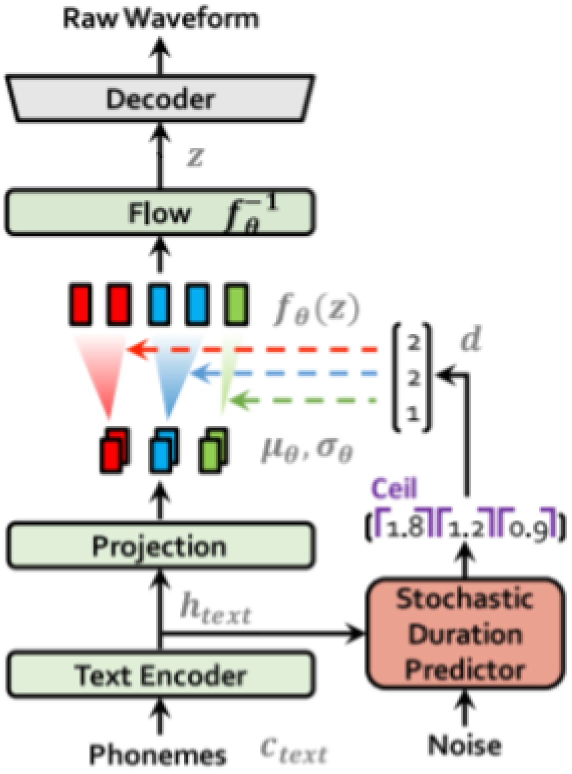


Fig 4. VITS Architecture: Illustrates the flow from phonemic input through text encoding, projection, and stochastic duration prediction, up to the generation of raw waveform output.

decoder, discriminator, and a stochastic duration predictor. The posterior encoder, built with non-causal WaveNet residual blocks, generates the normal posterior distribution mean and variance. The prior encoder includes a transformer-based text encoder and a normalizing flow that enhances the prior distribution's flexibility. The decoder, adopting the HiFi-GAN V1 generator design, uses transposed convolutions with a multi-receptive field fusion module. Discriminators are modeled after the multi-period discriminator architecture. The stochastic duration predictor employs residual blocks and neural spline flows to estimate phoneme duration, contributing to speech diversity and natural rhythm For this study, we utilized a VITS model pre-trained on the VCTK dataset[5], accessible via the Coqui TTS repository[6].

D. Face Blendshape Generation: Omniverse Audio2Face

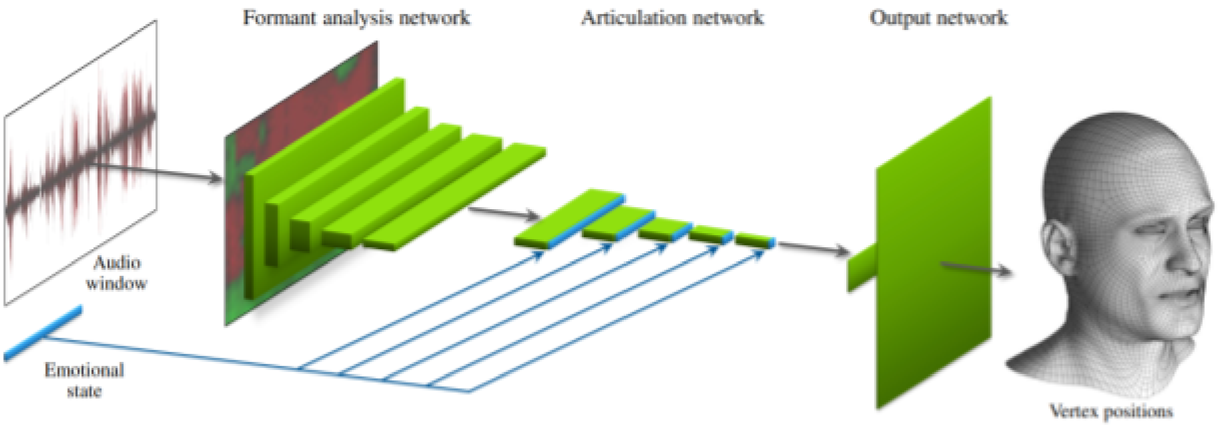


Fig 5. Nvidia Audio2Face Architecture: Illustrates the process from audio input through formant analysis and articulation networks, culminating in the generation of detailed facial blendshapes.

[5] https://datashare.ed.ac.uk/handle/10283/2651

[6] https://github.com/coqui-ai/TTS

In our system, the generation of facial blendshapes is a crucial component for creating lifelike avatar animations. Blendshapes are a technique in computer graphics where different facial expressions are formed by blending a set of predefined facial poses. These poses are each associated with specific facial muscle movements, enabling the creation of a wide array of human expressions [13].

For real-time and robust generation of these blendshapes, we employed NVIDIA's Omniverse Audio2Face [10]. This advanced facial animation framework uses generative AI to map audio inputs to facial animations, making it ideal for our interactive avatar needs. Audio2Face operates by animating 3D character models, such as the built-in “Digital Mark,” using a pre-trained deep neural network. This network translates audio cues into realistic facial animations, which are essential for enhancing the interaction experience between users and avatars.

The underlying neural network comprises a specialized layer, ten convolutional layers, and two fully-connected layers. It is structured into three sub-networks: the formant analysis network, the articulation network, and the emotional state representation as shown in Fig. 6. Initially, raw format data is extracted and refined, which is then processed to derive facial poses corresponding to the audio inputs. Concurrently, the emotional state representation augments the output at each layer of the articulation network to help differentiate between various facial expressions and speaking styles. The architecture employs strided convolutions to manage dimensions, with specific parameters chosen based on performance and training time considerations. Despite not retaining memory of past animation frames, the network demonstrates temporal stability in the animation output, indicating its robust capability in generating coherent and lifelike facial animations for interactive avatars.

We used 51 ARKit facial blendshapes standard[7] within this framework, where each blendshape represents a distinct facial muscle movement or expression. The combination of these blendshapes, driven by Audio2Face's real-time processing, allows for the creation of dynamic, nuanced facial animations that significantly enhance the realism of our avatars.

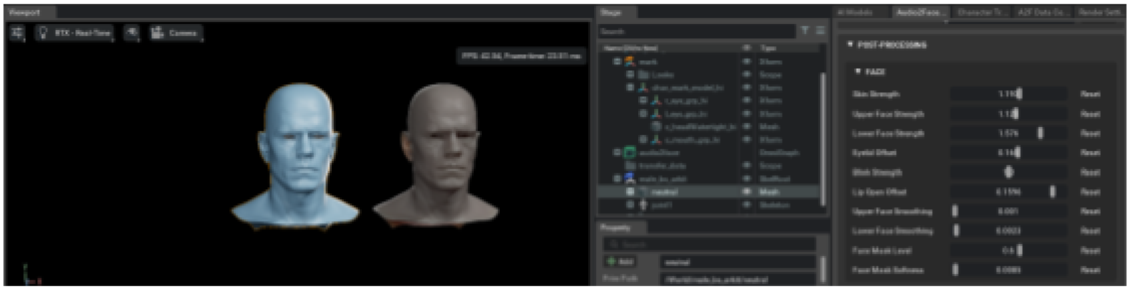


Fig 6. Audio2Face 2023.1.1

E. Avatar & Scene Rendering: Unreal Engine 5

Unreal Engine 5 (UE5) is employed for rendering the facial blendshapes created by NVIDIA's Omniverse Audio2Face. UE5's advanced rendering pipeline excels in real-time performance, essential for interactive systems. It skillfully handles lighting, shadowing, material application, and texture mapping, ensuring photorealistic quality of the facial animations.

The engine's sophisticated lighting and shadowing systems, enhanced by ray-tracing technology, add depth and realism to the scenes. This is crucial for rendering lifelike human facial expressions. Ray-tracing in UE5 simulates the physical behavior of light, producing highly realistic effects that contribute significantly to the natural appearance of the avatars.

[7] https://developer.apple.com/documentation/arkit/arfaceanchor/2928251-blendshapes

Additionally, UE5's material and texture management capabilities allow for detailed and realistic representations of human skin. This attention to detail in skin textures and colors is particularly important for achieving natural facial expressions in the avatars.

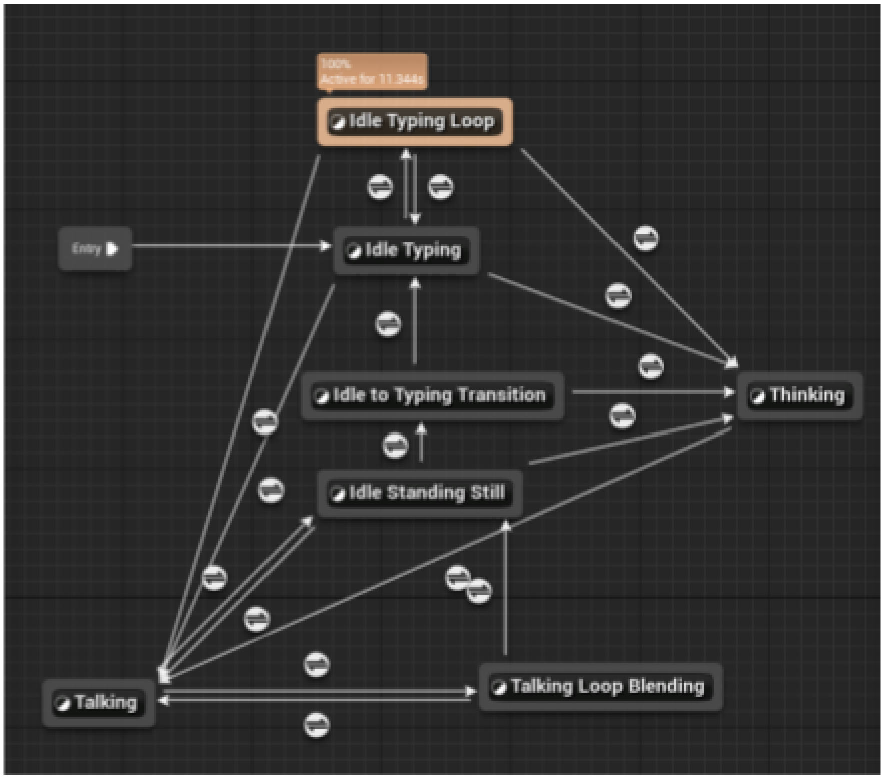


Fig 7. Avatar Animation State Machine in Unreal Engine

To achieve a lifelike representation of the avatar's demeanor, a state machine is implemented as the animation states of the avatar, as shown in Fig 8. This state machine comprises various states, including two idle states — 'idle typing' and 'idle standing still' — as well as 'talking' and 'thinking' states. Transition to the 'talking' state occurs upon the acquisition of facial blendshapes from the Omniverse Audio2Face system, which is indicative of active user engagement. This state is maintained for the duration of the interaction. Should there be a period of inactivity exceeding 30 seconds, characterized by the absence of new blendshapes, the avatar reverts to an idle state. During intervals between user input and the subsequent processing by the LLM, TTS, and Audio2Face, the avatar transitions into a 'thinking' state. This interim state serves to prevent any dissonance that might arise during processing delays, thereby preserving the fluidity and naturalness of the user-avatar interaction.

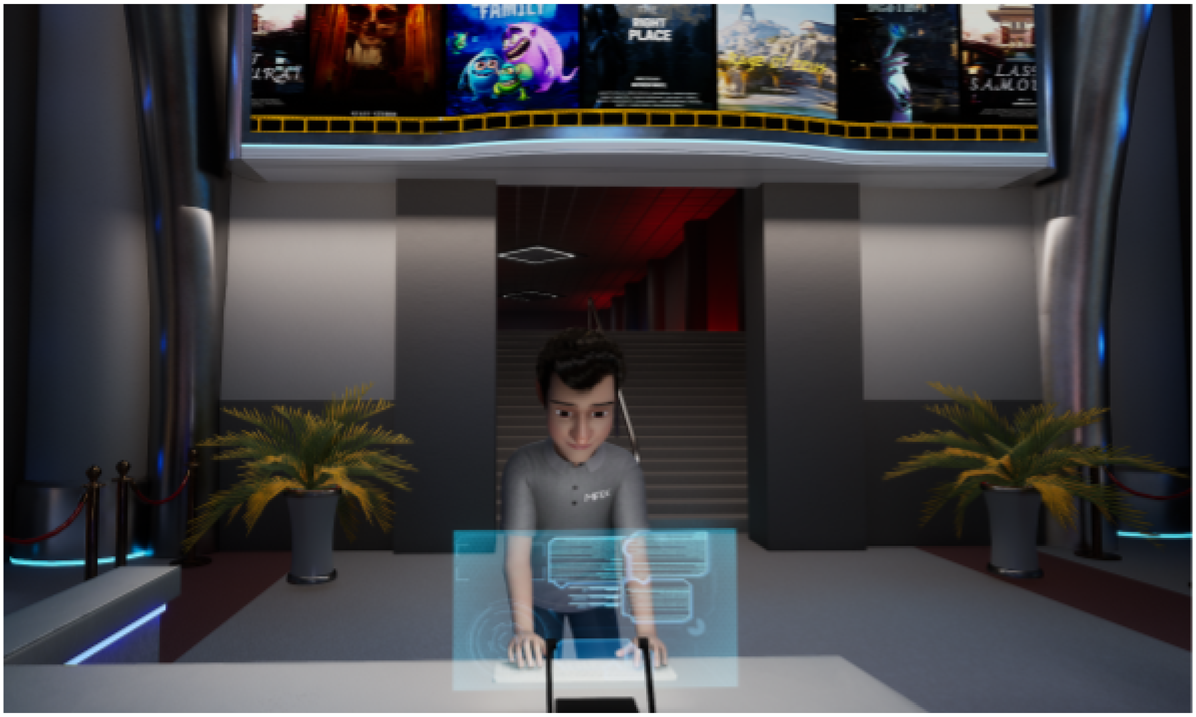


Fig 8. 3D Scene & Avatar in Unreal Engine 5

Fig 9 showcases the avatar in the context of the 3D-rendered environment, illustrating the system's interactive interface. The avatar, sourced from CGTrader[8], maintained a balance between a cartoonish aesthetic and high-fidelity detail. This design choice avoids the Uncanny Valley Effect while retaining a high polygon count to ensure visual quality. Such a

[8] https://www.cgtrader.com/3d-models/character/man/cartoon-man-rigged-70019e8b-0c8a-47f1-873e-c9678f2dfb47

representation is pivotal in fostering user engagement, as it combines approachability with the detailed realism necessary for an immersive conversational experience.

IV. Experiment

This experiment was conducted as a 2x2 between-subjects user study involving 53 participants between the ages of 18 to 34 (M = 22.8, SD=3.54), 16 of whom identified themselves as female. 87% of the participants completed the study at home. The study protocol number is 58-23-0450 and was approved by the university's Human Subjects Committee.

A. Design

This study was designed to assess the influence of visual representation and recommendation bias on user interaction with conversational agents. A set of controlled stimuli was developed to ensure consistency across the conditions and to address external variables potentially affecting the study's outcomes.

Four fictitious movie scenarios to serve as the basis for the experiment's stimuli. To achieve a balance in narrative engagement and to preclude participant bias due to familiarity with real-world media, the plots and titles were generated using GPT-4. This method ensured the creation of unique, yet equally compelling content while maintaining gender-neutral portrayals in protagonist names. Consistency in visual aesthetics was accomplished by producing poster images with Mid-Journey[9], utilizing standardized prompts to ensure each poster shared a similar stylistic approach. These measures aimed to

[9] Midjourney Documentation and User Guide

create a neutral baseline for participant interactions, focusing their evaluations on the conversational agents rather than the movie content itself. The details of these movies, including their titles, descriptions, and accompanying visual materials, are presented in Figure 5, illustrating the adherence to the design principles of uniformity and neutrality.

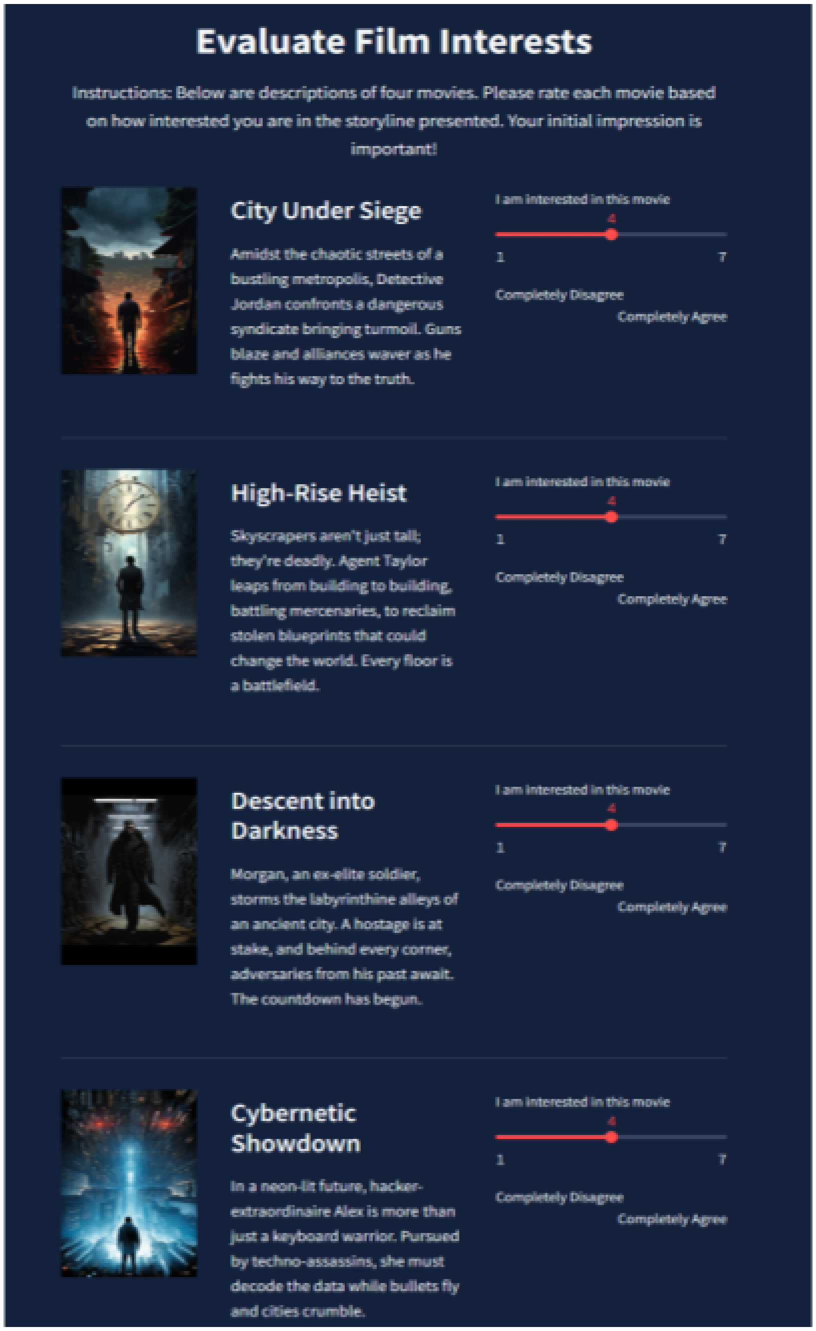
1. Visual Representation:

Participants engaged with either a lifelike human avatar or a simplified animated icon, both of which utilized the same underlying Large Language Model to ensure consistent verbal communication. The avatar was created using state-of-the-art animation to simulate realistic human interactions, aiming to explore the influence of visual fidelity on user engagement. In contrast, the icon offered a minimalistic but recognizable interface, akin to common digital assistants like Siri, to examine the effect of a more abstract representation.

2. Recommendation Bias:

The experimental design included manipulating recommendation bias by having the conversational agent exhibit a preference for one of the four movies. This was achieved by prompting the language model to mention the target movie more frequently and more favorably, thus introducing a subtle bias in its recommendations. The study's goal was to discern the impact of such bias on the participants' movie preferences and to measure this influence by observing changes in the ratings of the target movie following the interaction with the agent.

The utilization of a between-subjects design aimed at mitigating potential carryover effects that could obscure the interpretation of the results. By allocating different participants to each experimental condition, the design enhanced the study's ability to draw direct comparisons between the effects of the visual representation and recommendation bias. This approach was critical for isolating the variables of interest and ensuring that any detected differences in user perceptions were attributable to the intended manipulations within the conversational recommendation system (CRS). Understanding the extent to which recommendation bias can alter user opinions is essential in developing CRS that align with ethical standards and contribute to the field's knowledge base regarding user-agent interaction dynamics.



Evaluate Film Interests

Instructions: Below are descriptions of four movies. Please rate each movie based on how interested you are in the storyline presented. Your initial impression is important!

City Under Siege

Amidst the chaotic streets of a bustling metropolis, Detective Jordan confronts a dangerous syndicate bringing turmoil. Guns blaze and alliances waver as he fights his way to the truth.

I am interested in this movie

4

1 7

Completely Disagree

Completely Agree

High-Rise Heist

Skyscrapers aren't just tall; they're deadly. Agent Taylor leaps from building to building, battling mercenaries, to reclaim stolen blueprints that could change the world. Every floor is a battlefield.

I am interested in this movie

4

1 7

Completely Disagree

Completely Agree

Descent into Darkness

Morgan, an ex-elite soldier, storms the labyrinthine alleys of an ancient city. A hostage is at stake, and behind every corner, adversaries from his past await. The countdown has begun.

I am interested in this movie

4

1 7

Completely Disagree

Completely Agree

Cybernetic Showdown

In a neon-lit future, hacker-extraordinaire Alex is more than just a keyboard warrior. Pursued by techno-assassins, she must decode the data while bullets fly and cities crumble.

I am interested in this movie

4

1 7

Completely Disagree

Completely Agree

Fig 9. Overview of the Four movies Designed for Experiment

B. Procedure

The procedure began with participants entering a virtual meeting via Zoom, facilitated by the research conductor. After providing demographic details through a preliminary questionnaire, participants were presented with the movie scenarios. Initial interest levels were recorded on a 7-point Likert scale in response to "I am interested in this movie," setting a baseline for subsequent comparison.

Participants then engaged with a conversational agent, encountering one of the four experimental conditions. This phase allowed them to explore the movies further and adjust their interests based on the interaction. Participants had the freedom to end the dialogue at a point of their choosing, marking the natural end of their engagement. Upon the completion of the interaction, video and audio recordings were made, ensuring a detailed capture of the exchange for later analysis. The conversation transcripts, in particular, would serve as a critical resource for evaluating the conversational metrics defined in the study.

In the final phase, participants re-evaluated the movies, now informed by the interaction with the conversational agent. This post-interaction assessment aimed to measure any shifts in opinion resulting from the dialogue. The study was rounded off with a comprehensive post-study survey, where participants provided ratings and open-ended feedback on their experience, which was instrumental in gauging both the qualitative and quantitative aspects of user experience with the conversational agents.

C. Analysis

User experience is examined through subjective post-study survey ratings. Additionally, transcripts from each interaction are recorded and analyzed using the following quantitative and qualitative methods.

1. User Experience and Subjective Ratings

Participants rated their interest in movies on a 7-point Likert scale, both before and after the interaction, to assess shifts in their preferences. Additionally, engagement was inferred from the duration of the interaction, which participants could conclude at their discretion.

Post-study survey questions were also rated on a 7-point Likert scale, aimed at assessing the perceived naturalness and aesthetic appeal of the interaction with statements:

- “I found the movie recommendation bias in any way (4 being neutral).”
- “I found the interface visually pleasing (4 being neutral)”

The following metrics were quantitatively assessed, with the aim to employ Analysis of Variance (ANOVA) for identifying any statistically significant differences across the conditions:

Change in Target Movie Rating: The difference in Likert scale ratings of the target movie pre- and post-interaction, indicating the persuasiveness of the recommendation agent.

Total words spoken by Users: The aggregate count of words articulated by users during their interaction serves as an indicator of user engagement.

User Perceived Bias: This is evaluated through responses to the statement "I found the movie recommendation biased in any way (with 4 being neutral)," quantifying the user's perception of any bias present in the CRS.

Duration of Interaction: The length of time participants choose to interact with the agent, which may reflect the conversation's engaging nature.

Visual Appeal: Assessed through the response to “The interface was visually pleasing,” this metric evaluates the visual appeal satisfaction with the interface.

2. Quantitative Conversation Metrics

The transcript is analyzed based on two metrics: 1) frequency of movie mentions, 2) frequency of target movie mention.

Total Movie Mentions: This metric is determined by how often the movie is mentioned by either the recommendation agent or the user. This metric can be used to evaluate how engaging the interaction is.

Total Target Movie Mentions: This metric is determined by how often the target movie is mentioned by either the recommendation agent or the user. This metric can be used to evaluate how effective the bias recommendation is.

Total Words Spoken by Users: The aggregate count of words spoken by users during their interaction, serving as an indicator of user engagement.

V. Result

This section examines the results of the metrics from the previous section, showing the significance of the influence of the visual representation (Avatar vs. Icon) and recommendation bias (Bias vs. Unbiased).

A. User Experience and Subjective Ratings

Change in Target Movie Rating

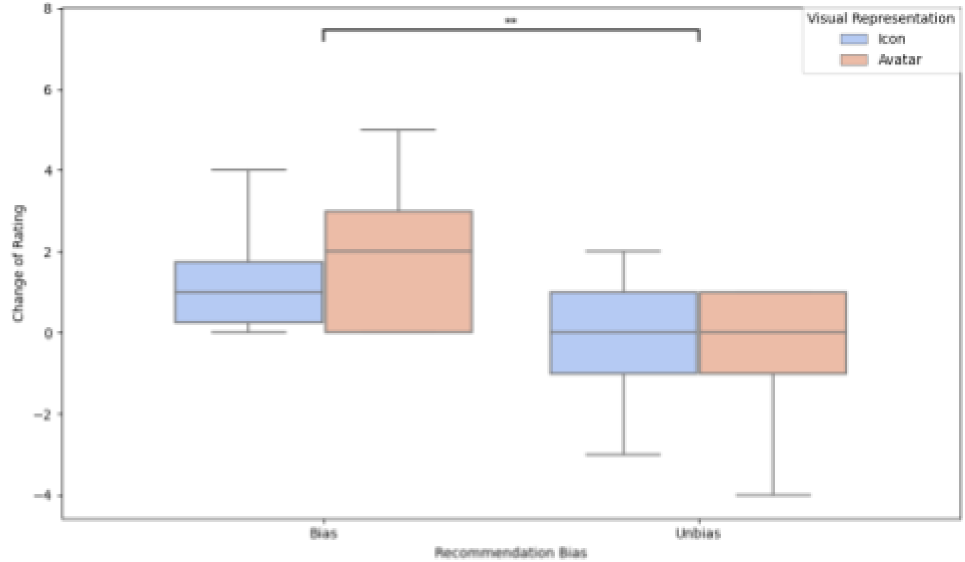


Fig. 10. Distribution of Rating Changes for the Target Movie by Recommendation Bias and Visual Representation

Figure 10 illustrates the distribution of rating changes for the target movie across different experimental conditions, specifically considering the influence of visual representation (Avatar vs. Icon) and recommendation bias (Bias vs. Unbias). The y-axis quantifies the change in user ratings, while the x-axis differentiates the conditions. Box plots represent the median value through the central mark and the interquartile range, detailing the spread of the middle 50% of the data.

The change in the target movie's ratings was not significantly affected by the type of visual representation ($F(1,49)=0.018$, $p=0.893$, $\eta^2<0.001$). However, recommendation bias had a significant effect on the change in ratings, as revealed by ANOVA ($F(1,49)=5.6298$, p=0.0021, η²=0.007). Subsequent post hoc analysis with Tukey's HSD test provided further insight, showing a mean difference in rating change of -1.7108 between the 'Bias' and 'Unbiased' conditions (95% CI [-2.5455, -0.8761], p=0.0001). This significant difference confirms the influence of recommendation bias on the change in movie ratings.

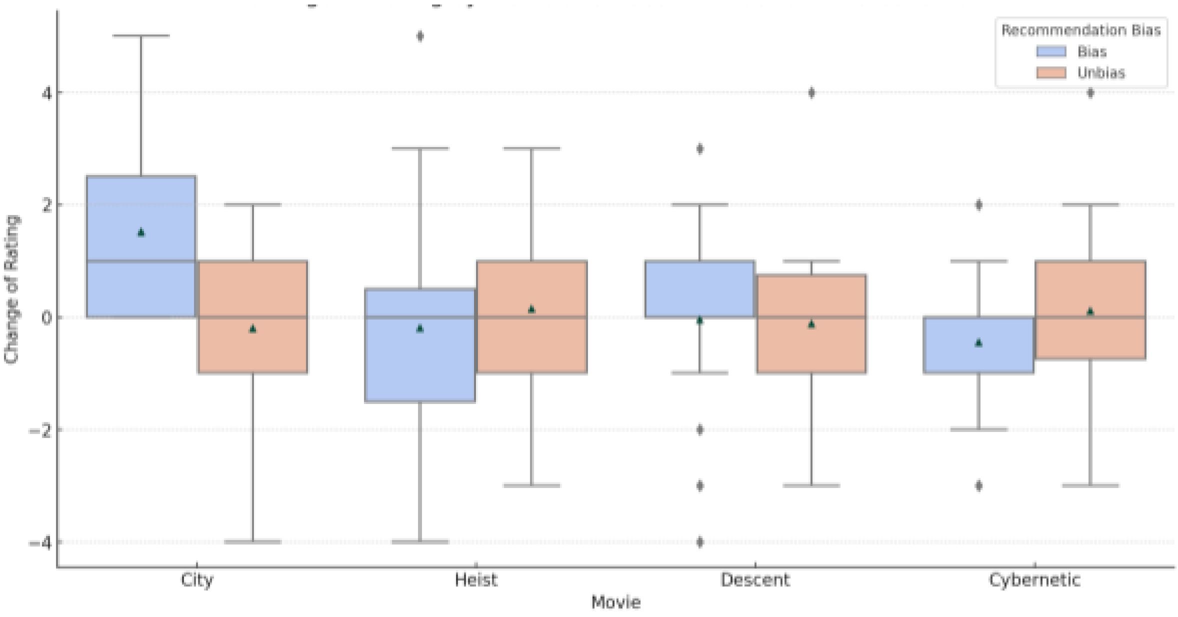


Fig 11. Distribution of Rating Changes for all movies by recommendation bias

Figure 11 presents a visual comparison of rating changes for all movies, distinguishing between biased and unbiased recommendation conditions. It shows that for the unbiased condition, the changes in ratings are relatively uniform across all movies, with no significant variations. This homogeneity suggests that when recommendations are unbiased, user rating changes are not influenced by the content of the movie. In contrast, under the bias condition, there is a visually discernible difference in the change of ratings for the target movie, indicating a greater influence of recommendation bias. This visual observation suggests that biased recommendations can lead to a more pronounced change in user ratings, particularly for the target movie, which stands out against the backdrop of other films. The subsequent ANOVA analysis will provide a statistical foundation for these observations, quantifying the influence of recommendation bias on rating changes.

The ANOVA conducted on the unbiased cases for the change in movie ratings showed no significant differences across the four movie categories ($F(3, 146) = 0.361$, $p = 0.781$, $\eta^2 < 0.001$). This lack of significant effect suggests that the type of movie did not influence the change in ratings when no recommendation bias was present. The subsequent post hoc analysis utilizing Tukey's HSD test aligned with the ANOVA results, indicating no significant mean differences in rating change among the movies. The largest observed mean difference was 0.5 between the 'Descent' and 'Heist' categories, but this was not statistically significant (95% CI [-0.3302, 1.3302], $p > 0.05$). The absence of significant differences suggests that the change in ratings is consistent across movies, assuming no recommendation bias.

Duration of Interaction

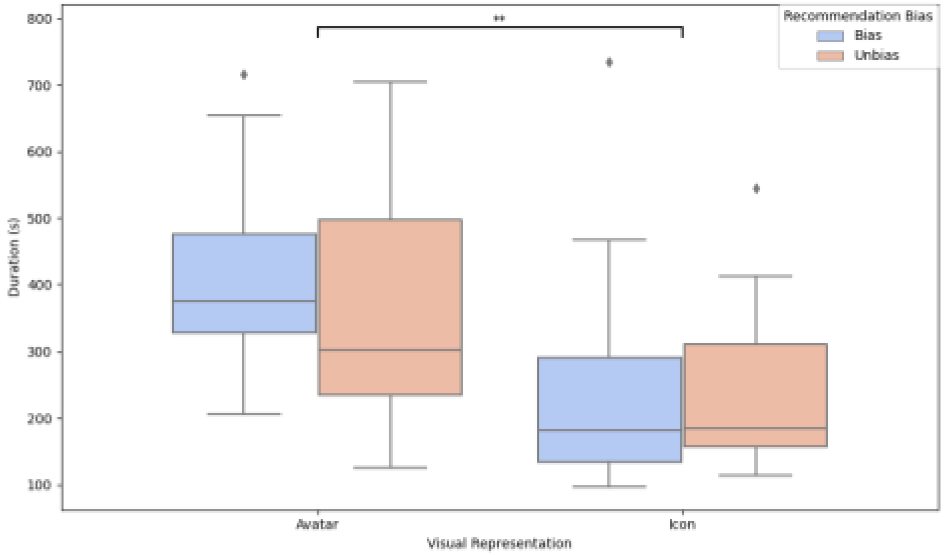


Fig 12. Distribution of Interaction Duration by Visual Representation and Recommendation Bias

Figure 12 illustrates the distribution of interaction durations in a Conversational Recommender System (CRS) across four experimental conditions that combine visual representation with recommendation bias. The conditions are split into Avatar with Bias, Avatar without Bias, Icon with Bias, and Icon without Bias. The box plots show the median, quartiles, and range of the durations.

The duration of interaction was significantly longer for the avatar compared to the icon, as evidenced by both ANOVA (F(1,49)=10.62, p=0.002, η^2=0.177) and the subsequent Tukey HSD test (mean difference = -147.32 seconds, p = 0.0018, 95% CI [-236.92, -57.73]). These results confirm that the type of visual representation has a significant effect on the duration, with avatars leading to longer interactions. However, there was no significant main effect of recommendation bias (F(1,49)=0.244, p=0.623, η²=0.004), nor was there a significant interaction between visual representation and recommendation bias (F(1,49) = 0.193, p=0.662, η²=0.003).

Perceived Bias

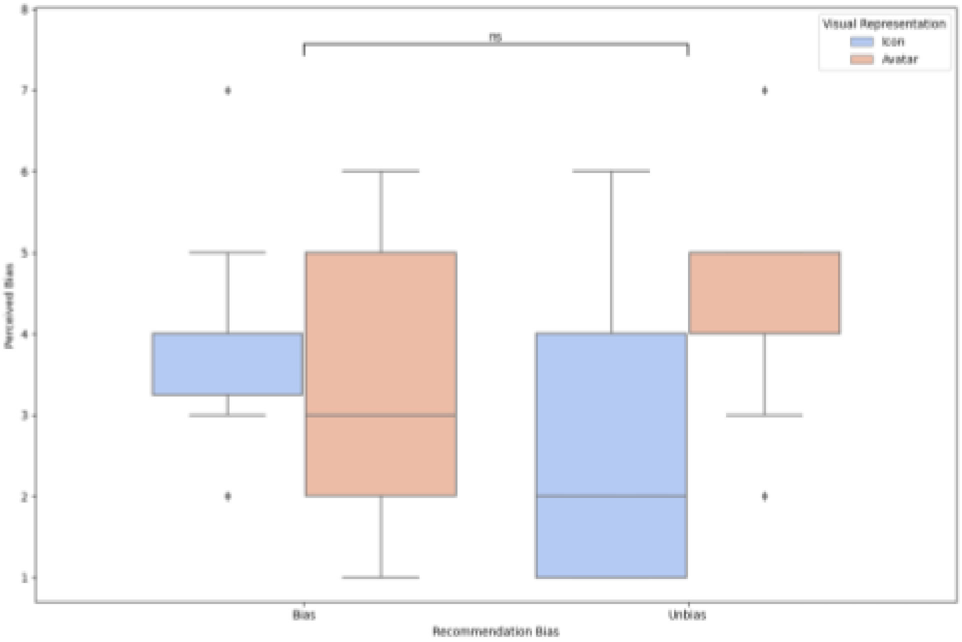


Fig. 13. User Perceptions of Biasness by Recommendation Bias.and Visual Representation

Figure 13 depicts the perceived bias among users in a Conversational Recommender System when exposed to either biased or unbiased recommendations, differentiated by the visual representation of either an icon or an avatar. The box plots represent the median, interquartile range, and full range of user-perceived bias, with no significant differences observed between the biased and unbiased conditions.

An independent sample t-test was conducted to compare the perceived bias between the biased and unbiased conditions. The results indicated no statistically significant difference in perceived bias between the two conditions (t(49)=0.717, p=0.476), suggesting that the manipulation of recommendation bias did not significantly alter users' perception of bias within the context of the CRS.

Visual Appeal

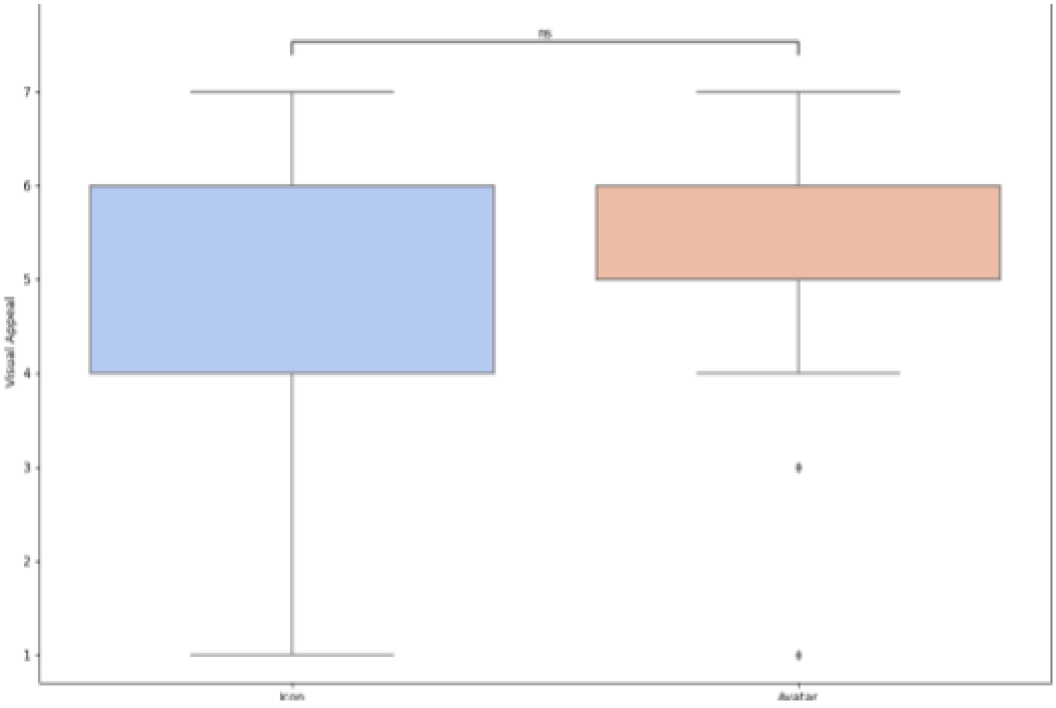


Fig. 14. User Perceived Visual Appeal by Visual Representation and Recommendation Bias.

Figure 14 displays user-reported perceptions of visual appeal related to visual representation in a Conversational Recommender System (CRS), differentiated by 'Avatar' and 'Icon' under conditions of 'Bias' and 'Unbias'. The y-axis represents the visual appeal scores, while the x-axis distinguishes between the visual representations. Box plots convey the median visual appeal scores, the interquartile ranges, and the span of responses, with outliers indicated by individual points.

ANOVA analysis revealed no significant main effects for visual representation (F(1, 49) = 0.764, p = 0.386) or recommendation bias (F(1, 49) = 1.904, p = 0.174), and no significant interaction effect between the two factors (F(1, 49) = 1.836, p = 0.182). This indicates that users’ feelings of visual appeal were not significantly affected by whether the visual representation was an 'Avatar' or an 'Icon', nor by the presence of recommendation bias. Tukey's HSD post hoc analysis supported these findings, showing no significant pairwise differences in the visual appeal ratings across the conditions. The results suggest that the visual appeal of the interface was perceived uniformly by participants, which is contrary to the expected uncanny valley effect often anticipated with human-like avatars. The analysis confirms that within the scope of this study, the design of the visual representation and the influence of recommendation bias did not lead to statistically different levels of perceived visual appeal.

B. Quantitative Conversation Metrics

Total Movie Mention

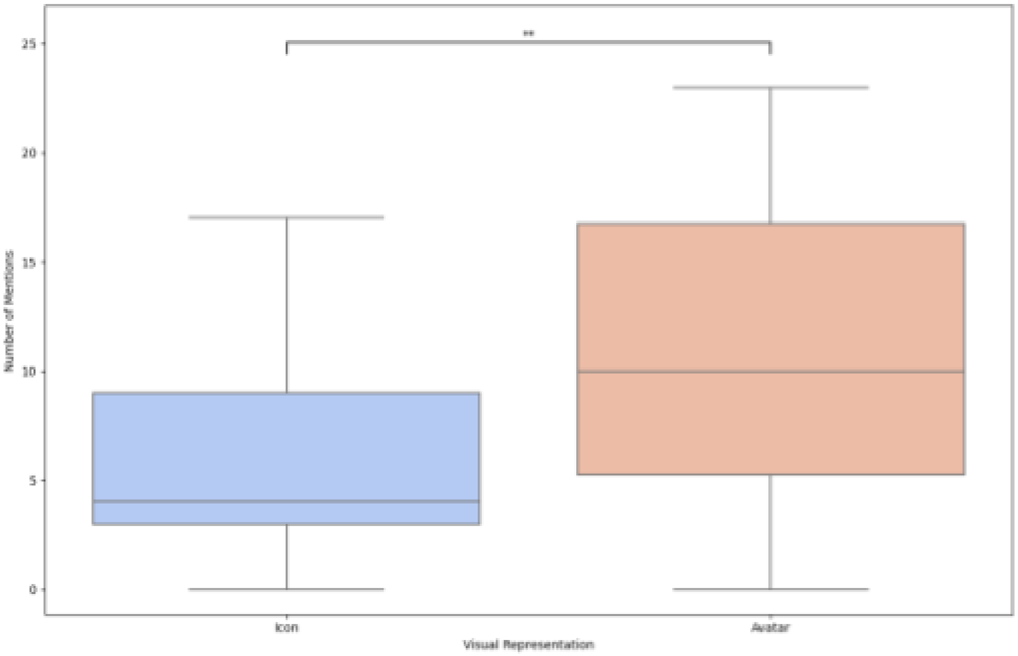


Fig 15. Total Movie Mentions by Visual Representation

Figure 15 shows distribution of movie mentions across two different visual representations within the experiment. The box plot illustrates a clear difference in the number of mentions between the simpler Icon and the more complex Avatar. Notably, the Avatar's interquartile range is broader and positioned higher on the y-axis, indicating more frequent mentions compared to the Icon. Outliers are marked, highlighting instances that deviate significantly from the typical mention count. This visual comparison suggests that users interacted more with the Avatar, an observation that is statistically substantiated by the subsequent ANOVA analysis.

ANOVA analysis revealed a significant effect of visual representation on the frequency of movie mentions (F(1,210)=18.759, p<0.0001, η^2=0.177), suggesting that different visual representations might influence engagement levels in movie discussions. Subsequent post hoc analysis with Tukey's HSD test confirmed this finding, showing that the Avatar visual representation leads to significantly more movie mentions than the Icon visual representation (mean difference = -1.1923, 95% CI [-1.735, -0.6496], p < 0.05). These results indicate a higher level of engagement with the Avatar compared to the Icon in discussions about movies.

Total Target Movie Mentions

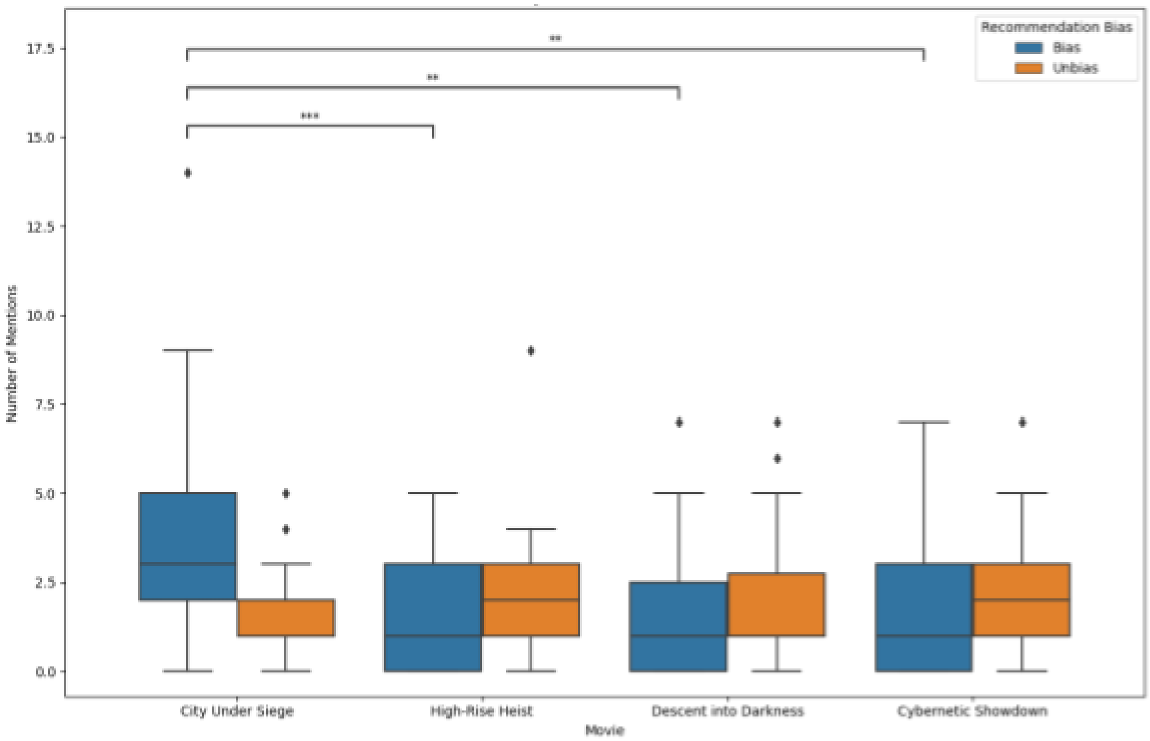


Fig 16. Amount of Movie Mentions by Recommendation Bias

Figure 16 illustrates a boxplot comparison of the amount of movie mentions. The x-axis categorizes mentions between the target movie "City Under Siege" and other movies, which include "High-Rise Heist", "Descent into Darkness", and "Cybernetic Showdown". The boxplot elucidates the distribution of mentions for the target movie, while the other movies' mentions are aggregated into individual box plots for a direct comparison. The hue distinction between biased and unbiased recommendations facilitates an evaluation of the impact that recommendation bias has on the frequency of mentions.

Statistical analysis using independent t-tests reveals a significant difference in the frequency of mentions between the target movie under biased conditions and other movies also under biased conditions. Specifically, "City Under Siege" received statistically more mentions than "High-Rise Heist" (p = 0.0007184), "Descent into Darkness" (p = 0.00125), and "Cybernetic Showdown" (p = 0.006349), confirming that the recommendation bias notably influenced the target movie's visibility within the CRS.

Total words spoken by Users

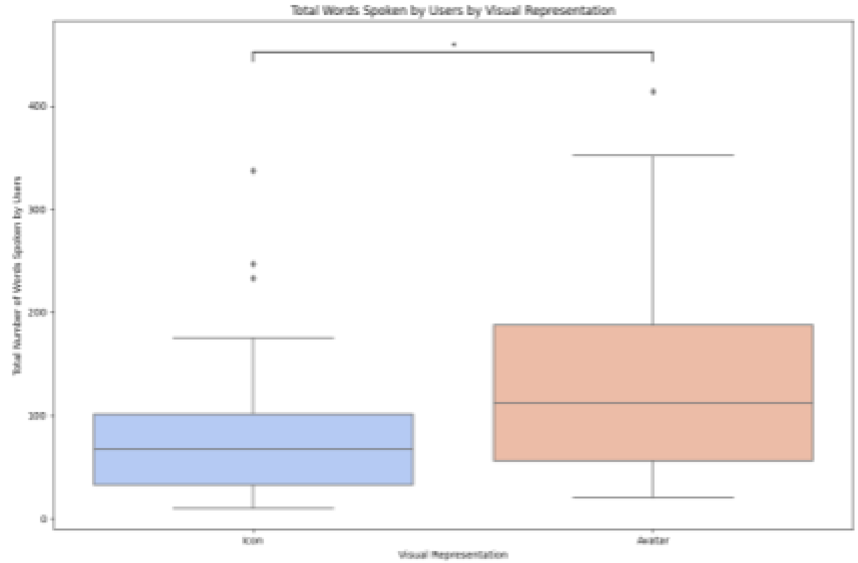


Fig 17. Distribution of User Word Count

Figure 17 depicts the total number of words spoken by users during the experiment under two different visual representations: an icon and an avatar. The boxplot illustrates the median, interquartile ranges, and outliers for word counts across these two conditions.

Statistical analysis was conducted using the Mann-Whitney-Wilcoxon test to evaluate the difference in word counts between the two visual representations. The test resulted in a p-value of 0.03023 and a U statistic of 245. These results indicate a statistically significant difference, suggesting that users engaged more when interacting with the avatar compared to the icon.

VI. Discussion

This section offers an analysis of the results obtained from the study while considering other factors that may have contributed to these outcomes.

A. Human Avatar vs. Animated Icon

The study's examination of visual representation in CRS reveals that user engagement is significantly heightened with the use of human-like avatars compared to animated icons. This is evidenced by prolonged interaction durations and increased user dialogue metrics, such as the total number of words spoken and movie mentions. These findings suggest that the visual complexity and human-likeness of avatars play a crucial role in capturing and maintaining user interest. Notably, this increased engagement does not directly translate to changes in users' subjective movie ratings. This distinction is critical, underscoring that while avatars can enhance the interactive experience and immersion, their influence on user preferences and decision-making is limited. This outcome prompts further investigation into how visual elements in CRS can be optimized to balance engagement and influence on user decisions.

In terms of visual appeal, this study’s design avoided the uncanny valley effect by employing avatars with a cartoonish aesthetic. This choice ensured that the avatars, while displaying human-like qualities, were stylized enough to avoid the discomfort and dissonance associated with hyper-realistic representations. The avatars' design fostered significant user engagement without eliciting the negative responses that the Uncanny Valley predicts. Consequently, our findings advocate for a balanced approach in avatar design within CRS, where stylization and human-likeness are harmonized to enrich user interaction, avoiding the uncanny valley's potential negative impact on user experience.

B. Bias Recommendation vs. Unbiased Recommendation

The introduction of recommendation bias into the CRS presents a complex interplay between system influence and user perception. Our study found that biased recommendations significantly shifted users' movie ratings. However, this bias was not overtly perceived by the users, indicating a discreet yet influential role of recommendation strategies on shaping user preferences. This raises important ethical considerations for CRS developers and designers, particularly in maintaining transparency and safeguarding user autonomy. The study highlights the need for responsible design approaches in CRS, ensuring that users are informed or aware of any underlying biases in recommendations. This aspect is vital in contexts where users rely on the impartiality of such systems for informed decision-making.

VII. Limitations and Future Work

One of the key constraints is the relatively simple task of movie recommendation, which was influenced by the limited reasoning ability and knowledge base of the open-source Large Language Model (LLM) employed. The rapid advancement in the field of open-source LLMs, however, opens up possibilities for exploring more complex tasks in future studies. As these models continue to evolve, offering enhanced reasoning capabilities and a more extensive knowledge base, they could enable the investigation of more intricate and nuanced user interaction scenarios. Such advancements would not only broaden the application scope of conversational recommender systems but also provide deeper insights into user behavior and decision-making processes in more complex domains.

Another area for further development is the system’s capacity for real-time representation of conversational agents. The current focus is primarily on facial blendshape generation, omitting the integration of body gestures. Efforts to incorporate Nvidia's Audio2Gesture[10] framework encountered challenges, particularly in the deformation of 3D models, suggesting a need for more sophisticated 3D modeling work. Achieving a comprehensive and immersive representation that includes both facial expressions and body gestures is crucial for enhancing the realism and interactivity of conversational agents. This aspect is particularly significant in creating more engaging and lifelike user experiences, essential for the efficacy and acceptance of conversational recommender systems.

VIII. Conclusion

This study's integration of Large Language Models with visual representations in Conversational Recommender Systems (CRS) unveils that interactive avatars, while significantly boosting user engagement, do not primarily drive recommendation persuasiveness. Instead, recommendation bias plays a more crucial role. The successful avoidance of the uncanny valley by the avatars suggests promising avenues for their wider application in user interfaces. These results contribute to the understanding of user interactions in CRS, emphasizing the need for balancing visual innovation with ethical considerations in recommendation strategies. Future research should focus on optimizing CRS designs by addressing the identified influence of recommendation biases and exploring the potential of ECAs in enhancing user experience in CRS environments.

[10] https://docs.omniverse.nvidia.com/extensions/latest/ext_audio2gesture.html

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." 2017.
2. Biao Zhang and Rico Sennrich. "Root mean square layer normalization." 2019.
3. Bickmore, Timothy W., et al. “A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology.” Journal of Biomedical Informatics, vol. 44,2, 2011, pp. 183-97. doi:10.1016/j.jbi.2010.12.006
4. Chat-REC: "Towards Interactive and Explainable LLMs-Augmented Recommender System."
5. Feng, Yue, et al. "A Large Language Model Enhanced Conversational Recommender System." arXiv preprint arXiv:2308.06212 (2023).
6. Frantar, Elias, et al. "Gptq: Accurate post-training quantization for generative pre-trained transformers." arXiv preprint arXiv:2210.17323 (2022).
7. Friedman, Luke, et al. "Leveraging Large Language Models in Conversational Recommender Systems." arXiv preprint arXiv:2305.07961 (2023).
8. Jannach, Dietmar, et al. "A survey on conversational recommender systems." ACM Computing Surveys (CSUR), vol. 54.5, 2021, pp. 1-36.
9. Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. "Roformer: Enhanced transformer with rotary position embedding." 2022.
10. Karras, Tero, et al. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." ACM Transactions on Graphics (TOG), vol. 36.4, 2017, pp. 1-12.
11. Kim, Jaehyeon, Jungil Kong, and Juhee Son. "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech." International Conference on Machine Learning. PMLR, 2021.
12. Korban, Matthew, and Xin Li. "A survey on applications of digital human avatars toward virtual co-presence." arXiv preprint arXiv:2201.04168 (2022).
13. Lewis, John P., et al. "Practice and theory of blendshape facial models." Eurographics (State of the Art Reports), vol. 1.8, 2014, pp. 2.
14. Li, Xian, et al. "Long Short-Term Planning for Conversational Recommendation Systems." International Conference on Neural Information Processing. Springer Nature Singapore, 2023.
15. Liu, Haiyang, et al. "Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
16. Lv, Zhihan, Fabio Poiesi, Qi Dong, Jaime Lloret, and Houbing Song. "Deep Learning for Intelligent Human–Computer Interaction." Applied Sciences, vol. 12, no. 22, 2022, https://doi.org/10.3390/app122211457
17. Mukherjee, Subhabrata, et al. "Orca: Progressive learning from complex explanation traces of gpt-4." arXiv preprint arXiv:2306.02707 (2023).
18. Noam Shazeer. "Glu variants improve transformer." 2020.
19. Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." International Conference on Machine Learning. PMLR, 2023.

20. Seyama, Jun'ichiro, and Ruth S. Nagayama. "The uncanny valley: Effect of realism on the impression of artificial human faces." Presence 16.4 (2007): 337-351.
21. Sun, Yueming, and Yi Zhang. "Conversational recommender system." The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018.
22. Timothy W. Bickmore, Daniel Schulman, and Candace L. Sidner. "A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology." Journal of Biomedical Informatics, vol. 44, 2, 2011, pp. 183–197.
23. Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).
24. van Pinxteren, Michelle M.E., et al. "Effects of communication style on relational outcomes in interactions between customers and embodied conversational agents." Psychology & Marketing (2023).
25. Wang, Xiaolei, et al. "Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models." arXiv preprint arXiv:2305.13112 (2023).
26. Wu, Likang, et al. "A Survey on Large Language Models for Recommendation." arXiv preprint arXiv:2305.19860 (2023).
27. Zhang, Biao, and Rico Sennrich. "Root mean square layer normalization." 2019.
28. Zhu, Jiarui, et al. "Free-form Conversation with Human and Symbolic Avatars in Mixed Reality." 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE Computer Society, 2023.