Navigating with FurGuide: The Impact of Narratives on Instruction-Following Tasks

Shivaanee Eswaran, Omar Riyaz,
Yuting Hang, Ziqi Chen, Jiayi Wang, Jiahao Zhang,
Zhiqing Wang, Xiaokun Yin, Bruce Wilson, Javier Chiyah-Garcia
Heriot Watt University

Abstract

This study investigates how user experience and recall during instruction-following tasks in a virtual environment are influenced by narrative-based interaction with FurGuide, an expressive social robot. Users interact with FurGuide, which provides them with instructions on navigation and tasks, either with or without narrative support, as they go through a Minecraft map. To evaluate the impact on user recall, task performance, and engagement, we analyze data on navigation and task accuracy, the number and type of follow-up questions, and responses to questionnaires filled out after the experiment. The results suggest that FurGuide's likability and perceived anthropomorphism are improved by narratives. While other differences between the narrative and non-narrative experiment conditions are minimal, the narrative condition exhibited slightly better outcomes.

1 Introduction

The field of human-robot interaction has seen the emergence of creative techniques for user help and guiding due to the incorporation of powerful large language models (LLM). LLMs combined with robotic systems have brought about a paradigm shift in human-robot interaction by providing unmatched natural language understanding and task execution capabilities (Zhang et al., 2023).

FurGuide [1] is an example of this combination in action, using OpenAI's GPT-3.5 Turbo model [2] to give users specific assistance in completing tasks navigating their way around in a Minecraft environment. FurGuide uses natural language processing (NLP) to engage users in real-time conversation, offering detailed instructions and contextual guidance to help them successfully complete their tasks and navigate through all the rooms in the Minecraft map.

[1] A demo of our system can be found here
[2] GPT-3.5 Turbo Docs

Figure 1: A user interacting with the FurGuide while navigating through the virtual environment.

The impact of narrative-based interaction with respect to FurGuide's guidance system is investigated in this study. Our experiment aims to assess how well story-telling elements improve user recollection, experience, and engagement using two variations of the robot, one that uses narratives and the other that does not. To be more precise, we aim to answer the following research questions:

- RQ1: How do the use of narratives during instruction-following tasks affect the recall, usefulness, and accuracy of the instruction set?
- RQ2: How do the use of narratives during instruction-following tasks affect the number and type of follow-up questions?
- RQ3: How does the use of narratives affect the likeability and anthropomorphism of the Furhat robot?

2 Literature Review

2.1 Narratives

In human communication and cognition, narratives are crucial because they influence how people perceive, understand, and retain information. In the context of human communication and interaction, narratives are made up of elements of structured storytelling that are intended to provoke certain

feelings in the audience, engage them in a meaningful and coherent way, and convey information. Narratives in our experiment refer to structured storytelling elements integrated into interactions with the Furhat robot.

The results in (Araujo and Kollat, 2018) highlight how crucial it is to have engaging strategies and storytelling elements in Corporate social responsibility (CSR) messaging. According to the study, brands and businesses on Twitter that tweeted about CSR with more usage of emotions and aspirational talk were linked to better levels of endorsement and content diffusion.

The findings from (Kromka and Goodboy, 2019) showed that students in the narrative lecture condition thought more highly of the teacher and expressed a desire to take another course with them in the future. Additionally, compared to students in the examples condition, students in the narrative condition reported paying closer attention to the lecture for longer periods of time and did marginally better on a short-term recall test.

In (Wolfe and Mienko, 2007) the results show that individuals with average backgrounds learnt just as much from narrative and informational literature. However, informational texts were more likely to be remembered by individuals with extensive prior knowledge of the subject matter than narrative ones. It's interesting to note that memory retention of the narrative material was mostly unaffected by prior knowledge. The study indicates that narrative and informational texts function differently when it comes to fusing new knowledge with what we already know.

In (Ulsamer et al., 2019) it talks about an experiment where participants had to navigate a virtual campus. A Virtual Reality (VR) video with narratives was viewed by some people, whereas a VR video without narratives was watched by others. Those who had the storytelling experience had a better memory for important stops along the way. The paper shows that using storytelling elements to connect emotionally charged video segments helps them become more remembered. Compared to users without storytelling cues, those with them navigated faster and with fewer steps.

2.2 The Furhat Robot

The Furhat Robot is a social robot developed by the Swedish company Furhat Robotics [3] which is used in customer service and education, that has a lifelike human face on a screen and can interact with people in conversation and convey emotions (Al Moubayed et al., 2014). The Furhat is a robot head made up of an animated face that is projected onto a 3D mask that is designed to mimic the animated face on it using a small projector (Al Moubayed et al., 2013). It is equipped with a pan-tilted neck and a 3D face display, ensuring accurate transmission of gaze and orienteering movement. The Furhat makes use of advanced dialogue toolkits that are meant to enable multimodal, multiparty, spoken, and located human-machine dialogue (Al Moubayed et al., 2013).

The Furhat robot was selected as the platform for the study due to its distinct features and suitability for enabling authentic human-robot interactions. Additionally, it is a great option for meeting the requirements of our study due to its flexibility and adaptability. Because of its modular design, it may be easily integrated with a wide range of software components, including the GPT-3.5 model from OpenAI. Facial expressions, domain-specific information, and prompt-engineered domain-general discussions are smoothly combined by the Furhat robot and OpenAI's GPT 3.5 (Cherakara et al., 2023). By using a large language model, The Furhat can respond with precision and detail, keep the flow of the discussion flowing, and use dynamically managed non-verbal cues in encounters. (Hanschmann et al., 2024).

3 System Design and Architecture

Figure 2 shows the system architecture, which includes a conversational framework that allows users to communicate verbally with FurGuide. FurGuide's system architecture consists of multiple essential parts, such as Automatic Speech Recognition to convert user speech into text, Natural Language Understanding to interpret that text, a Dialogue manager and Natural Language Generation powered by GPT-3.5 Turbo to generate responses that sound genuine. This, along with the prompt engineering, the virtual environment, and the Furhat buttons, makes up our entire system for the experiment.

3.1 Automatic Speech Recognition

Automatic Speech Recognition (ASR) has long been seen as an essential link for encouraging improved communication between humans and be-

[3] https://furhatrobotics.com/

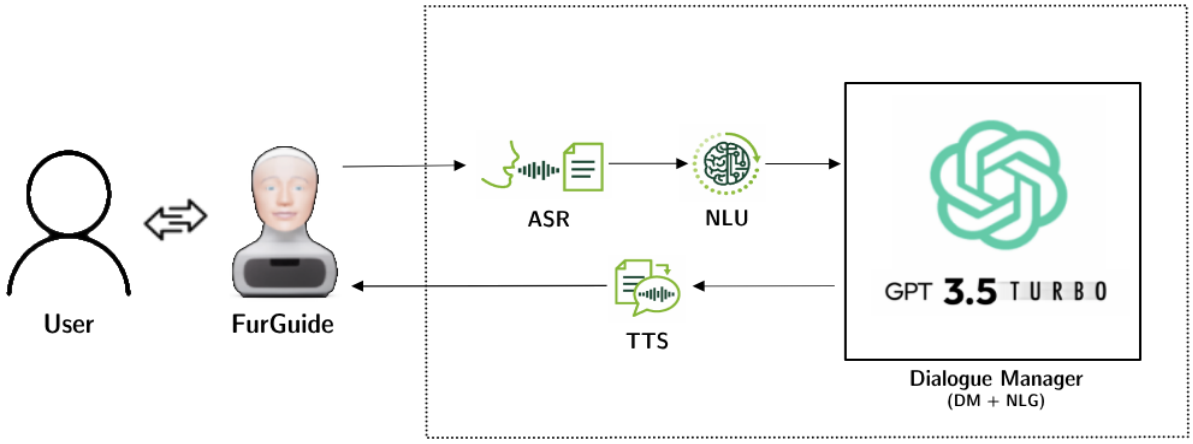

Figure 2: FurGuide's System Architecture.

tween humans and machines (Yu and Deng, 2016). The Google Cloud Speech-to-Text module [4] is used by FurGuide's system to do ASR. This module, which is automatically integrated into the system via the Furhat SDK, uses machine learning techniques to translate spoken words into text.

3.2 Natural Language Understanding

Since its inception in the 1950s, computer-based Natural Language Understanding (NLU) continues to pose a significant challenge in terms of allowing computers to understand the intricacies of human language (Bates, 1995). The Furhat SDK's integrated dialogue manager controls dialogue states and preserves conversational flow in line with intents recognized by the NLU component.

3.3 Dialogue Management

Dialogue Management consists of two submodules: Dialogue Manager (DM) and Natural Language Generation (NLG). OpenAI's GPT-3.5 Turbo, a high-performance large language model known for its powerful natural language understanding and processing abilities, powers these modules. Our system uses the gpt-3.5-turbo-0125 model, which is one of the most powerful models in the 3.5 series, and it is priced at 0.0005 dollars per 1000 tokens. The NLG module creates natural-sounding responses based on input given, while the DM maintains dialogue states and interactions.

3.4 Prompt Engineering

Even the most advanced models have quirks when it comes to language generation. One notable issue is 'hallucination'. This occurs when LLMs add on some false information by mistake. The work in (Xu et al., 2024) tells us that LLMs are unable to learn every computable function, thus they will forever experience hallucinations. From the findings in (Amatriain, 2024), we know that hallucination rates can be considerably reduced by using prompt engineering to establish limits of acceptable responses and directing the LLM toward more accurate outputs.

```
initialPrompt = "You are FurGuide, an extroverted navigational assistant. You are helping a
 user complete the specified task then navigate through a door in a virtual environment. The
  room the user is in is described below, the task the user must do in that room is also
 described below. The user must complete the task in the room and then navigate to the next
 room by going through a door. The navigational instructions are given with narratives. If
 they user asks about what they need to do, tell them what the navigational instructions
 are, but leave out the narrative element, focus only on the navigation. You provide
 guidance to complete the task and navigational instructions with the given narratives, and
 answer any doubts or queries the user may have regarding these. If the user asks a question
  that cannot be answered with the information given below, do not make up an answer,
 instead tell the user that you cannot provide that information to them. You answer
 questions in a extroverted way." +
```

Figure 3: An example of the initial prompt given to the LLM.

We used prompt engineering to provide the GPT-3.5 Turbo model with precise instructions and context, turning its responses into FurGuide, the navigational assistant. Figure 3 shows an example of an initial prompt for one of the rooms in the map. FurGuide was told by the prompt to guide users through each room's tasks, provide navigational instructions for navigating the virtual environment, and answer questions from users about these tasks and instructions. The model received detailed room descriptions, corresponding tasks, and navigational instructions. FurGuide was able to ensure a smooth navigational experience by using this information to accurately reply to user follow-up questions and provide relevant replies depending on the context of the room the user was currently in.

[4] https://cloud.google.com/speech-to-text

3.5 Text To Speech

FurGuide uses the Amazon Polly [5] service, which is already integrated with FurhatOS, to transform text to speech. (Cambre et al., 2020) offers advice on choosing text-to-speech (TTS) voices for lengthy content. Considerations including voice clarity, quality, and overall listening experience should be made when selecting a voice. The results from (Dodd et al., 2023) show that American English TTS voices were rated as less human-like when compared to British English and Indian English TTS voices. Additionally, findings from (Crowelly et al., 2009) suggest that perceived gender characteristics in robot voices can impact human-robot interactions. Subjects from this study found the male voice more reliable for the robot; thus, in our experiment, we incorporated the 'Brian' voice from Amazon Polly, which is a British English male voice.

3.6 The Virtual Environment

The virtual environment created in Minecraft [6], which mimics a maze inside a house, serves as the setting for the tasks and navigation participants have to complete. Navigating through mazes in Minecraft simulates the ambiguous and confusing nature of real-world situations (Adi, 2016). There are two distinct routes that each have four rooms, with a task to complete in each room. The rooms give participants three doorways to exit through, where each door opens to a hallway with another door on the other end, opening into the next room. Figure 4 shows one of the rooms created in Minecraft. You can see the door leading to the white hallway in this figure.

Only one door is the correct one, which leads to the next room, but the map was designed in such a way that whichever door the participants go through, they are sent to the hallway corresponding with the correct door using the Minecraft 'teleport' command. This enables us to hide any errors the participants make from them while we, on our end, can see and take note of them. Participant guidance through these tasks is greatly enhanced by instructions, which correspond with methods of problem-solving found in the actual world (Zhao et al., 2023). In order to successfully navigate through both mazes, participants must rely on their navigational abilities and guidance from FurGuide.

[5] https://docs.aws.amazon.com/polly/

[6] https://www.minecraft.net/en-us

Figure 4: The final room for the first route.

3.7 The Wizarding Interface

The Furhat robotics platform offers developers many human-computer interaction tools, including the 'Wizarding' technology for button capabilities (Al Moubayed et al., 2015). This approach makes it possible to precisely control activities through manual manipulation, making it easier to demonstrate experiments.

We use these buttons to first choose what route they take and which version of FurGuide helps them throughout the route. For example, one participant could start with the FurGuide with narratives implemented for the second route and then move onto the FurGuide without narratives for the first route. Then, within a route, there are buttons for all four rooms of the route, a pause button, and an end route button for when the participant completes the route. This is so that FurGuide is aware of which room the participant is in currently and can answer any questions they may have about the room or the tasks in the room. When a participant enters a room, on our end in the Furhat Dashboard, we press the button corresponding with the room they enter into.

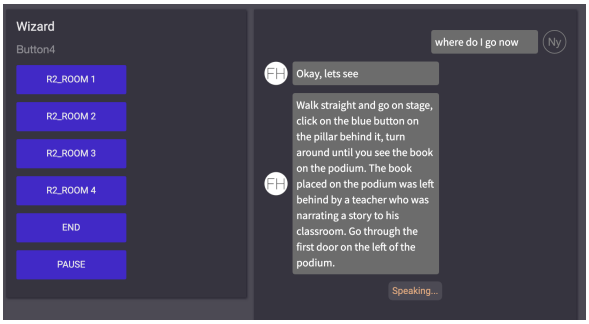

Figure 5: The buttons on the Furhat web interface dashboard for our experiment, with an example dialogue next to it.

4 Implementation

The experiment was implemented using an elaborate setup that included two laptops, a keyboard, mouse, monitor, the FurGuide robot, and a microphone. The setup for the experiment is shown in Figure 6, where one laptop (on the right) ran Minecraft in addition to Microsoft Forms for gathering participant data. This laptop's screen is mirrored onto the monitor by connecting it via a USB-C cable. The second laptop (on the left) ran the FurGuide Kotlin code and had the Wizarding Interface open on the Furhat Dashboard.

Figure 6: The experiment setup

Participants were seated at the monitor at the start of the experiment, and FurGuide was introduced to them along with a briefing about the study. After a demonstration of how participants might have conversations with FurGuide, the structure and procedures of the experiment were explained. Using the connected keyboard and mouse, participants then filled out the consent form and pre-experiment questionnaire.

Participants were then familiarized with the basic controls of Minecraft before starting the navigation. Based on predefined routes and FurGuide variations (narrative or non-narrative) assigned according to participant IDs, FurGuide would offer guidance. Then, on the second laptop, the FurGuide code was run, and a button was clicked on the Wizarding Interface for route selection and FurGuide variations.

Figure 7 shows the setup of the Wizarding Interface. Through the use of the Wizarding Interface, a within-subjects design was implemented. Within-subject designs account for individual variations. By acting as their own control, each participant helps to reduce the impact of differences in personal traits between conditions (Simkus, 2023). So in our experiment, if a participant starts off on route

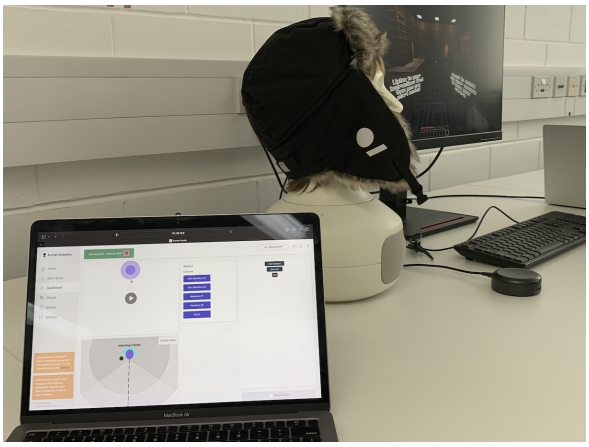

Figure 7: The Wizarding Interface Setup

2 and the narrative variation of the FurGuide, the next route they take will be route 1 without narratives on the FurGuide. This order of routes and variations was randomized to reduce order effects and potential biases.

Using OBS [7] on the first laptop, we start to record the participants screen. FurGuide then provides its instructions in accordance with the pressed button. After hearing all of the instructions, the player will begin navigating around in Minecraft. FurGuide provided guidance and answered participant questions about tasks, navigation, and Minecraft controls during the navigation. The Wizarding Interface now shows buttons for each room of the route. The FurGuide's responses were triggered by button presses corresponding to participant progress through the virtual environment. Finally, the 'end' button is then pressed, informing the user that the route has been finished.

We monitor the participant's progress while they navigate in Minecraft by keeping an eye on participant performance, recording task and navigation errors, frequency and type of follow-up questions, and completion time. The first laptop's screen display was used for observations, which made it easier to gather these metrics.

After finishing a route, participants returned to Microsoft Forms to complete the post-task questionnaire for that route. The procedure was then repeated for the next route, maintaining consistency in the experiment with the predetermined route order. Once both routes and post-task questionnaires were completed, the screen recording would stop, and the participant would be informed that they were finished with the experiment.

[7] https://obsproject.com/

	Recall-NN	Recall-N	Usefulness-NN	Usefulness-N	Accuracy-NN	Accuracy-N
Mean	2.872	3.070	6.167	6.167	0.708	0.458
Std	1.853	2.231	0.816	0.761	0.690	0.721
Median	3.083	2.917	6.000	6.000	1.000	0.000
Min	0.000	0.000	4.000	5.000	0.000	0.000
Max	6.350	7.400	7.000	7.000	2.000	2.000

Table 1: Results from RQ1, where NN stands for Non-Narratives group and N stands for Narratives Group

	Questions-NN	Questions-N	Likea-NN	Like-N	Anthro-NN	Anthro-N
Mean	1.875	1.667	4.288	4.424	3.500	3.125
Std	1.028	1.119	0.708	0.750	1.022	1.676
Median	1.875	1.500	4.100	4.900	3.500	3.500
Min	0.25	0.000	3.000	2.800	1.000	5.000
Max	3.4	4.500	5.000	5.000	5.000	7.000

Table 2: Results from RQ2 and RQ3, where NN stands for Non-Narratives group and N stands for Narratives Group

5 Evaluation

5.1 Metrics

The objective and subjective measures collected include:

- **Pre-experiment:** Demographic information about participants.
- **Each Experiment Condition:** Instruction recall, usefulness (Almere model), animacy and anthropomorphism (Godspeed sub-scale)
- **Experiment Administrator:** Time taken to complete each experiment condition, type and number of participants' follow-up queries, and navigation details when instructions are incorrectly followed.

The instructions recalled is evaluated with a gist-based memory recall method by comparing it with a predetermined keywords list of each instruction, to produce recall scores for each participant for both conditions. To then test each research question, a Student's t-test is performed. This statistical test is performed with the recall scores, usefulness ratings, and navigation details collected by the experiment administrator to test **RQ1**. To test **RQ2**, the t-test is performed with the type and number of follow-up questions, also collected by the administrator. **RQ3** is evaluated using t-values for the animacy and anthropomorphism measures.

Statistical significance is accepted when t-value > 1.3 (the critical value for a two-tailed test with alpha=0.20 and 46 (sum of sample sizes - 2) degrees of freedom) and p-value <= 0.05.

	t-Value	p-Value
Recall	0.334	0.740
Usefulness	0.398	0.696
Accuracy	-0.20	0.860
Follow-Up	-0.215	0.837
Likeability	2.997	0.017
Anthropomorphism	2.340	0.047

Table 3: Student's t-tests for evaluating RQs

5.2 RQ1 Results

Measuring the accuracy of the instruction set recalled was done in two ways. Firstly, the instructions are divided into task and navigation instructions. Then, the percentage of keywords successfully recalled by the participant for each set of instructions was added up. In the second method, a point was assigned for each instruction set that was recalled perfectly. This point was awarded if the participant recalled all keywords for the instruction; otherwise, no point was given.

The t-tests performed to evaluate **RQ1** showed no statistically significant results. The **perfect navigation score** measure showed marginal statistical significance. This means that the perfect navigation score was higher for the *'with narrative'* experiment condition.

5.3 RQ2 Results

Similar to the previous evaluation, no statistically significant results were derived to answer **RQ2**.

5.3.1 RQ3 Results

Student's t-tests for the likability (t-value = 2.997, p-value = 0.017) and anthropomorphism (t-value = 2.340, p-value = 0.047) measures yielded significant results. This answers **RQ3** with the narrative condition being more liked and higher perceived anthropomorphism scores.

6 Conclusion and Future Work

In conclusion, our study investigated how user navigation tasks in a virtual environment were affected by narrative-based interaction with a Furhat robot. We investigated how narratives affected user recall, usefulness, and accuracy during instruction-following tasks through careful design and implementation. We also measured how narratives affect the number and type of follow-up questions and how they affect the likeability and anthropomorphism of the robot. Although we arrived at statistically insignificant results for **RQ1** and **RQ2**, from the results of **RQ3**, we can conclude that the narrative experiment condition is better liked and perceived more anthropomorphically.

However, there were a number of limitations to our study that should be taken into account. First off, our results might not have been as reliable as they could have been if the sample size had been larger. Our experiment required 28 participants to provide statistically significant results. This was found using a statistical power analysis tool called G*Power [8] (Faul et al., 2009). Our experiment only had 24 participants. Furthermore, the majority of participants were international students, which may have introduced language proficiency differences. These differences resulted in misunderstandings and variations in response accuracy.

Future studies may attempt to overcome these limitations and deepen our knowledge of narrative-based interactions in human-robot communication. Findings can be made more reliable and broadly applicable by expanding the number of participants to include people from different demographic backgrounds.

As a result, even though our study provides insightful information on the function of narratives in navigation tasks, more investigation is necessary to fully understand these dynamics and progress in the field of human-agent interaction.

[8] gpower docs

Acknowledgements

The authors would like to thank the professors of the F20CA class at Heriot-Watt University as well as the other group members for this class that participated for our experiment.

References

Adi. 2016. Researchers develop AIs to tackle Minecraft mazes — adigaskell.org. https://adigaskell.org/. [Accessed 21-04-2024].

Samer Al Moubayed, Jonas Beskow, and Gabriel Skantze. 2013. The furhat social companion talking head. In *INTERSPEECH*, pages 747–749.

Samer Al Moubayed, Jonas Beskow, and Gabriel Skantze. 2014. The world's most advanced social robot - Furhat Robotics — furhatrobotics.com. https://furhatrobotics.com/. [Accessed 20-04-2024].

Samer Al Moubayed, Jonas Beskow, and Gabriel Skantze. 2015. Furhat robot and developer documentation. https://docs.furhat.io/. [Accessed 21-04-2024].

Xavier Amatriain. 2024. Measuring and mitigating hallucinations in large language models: Amultifaceted approach.

Theo Araujo and Jana Kollat. 2018. Communicating effectively about csr on twitter: The power of engaging strategies and storytelling elements. *Internet Research*, 28(2):419–431.

Madeleine Bates. 1995. Models of natural language understanding. *Proceedings of the National Academy of Sciences*, 92(22):9977–9982.

Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Neeraj Cherakara, Finny Varghese, Sheena Shabana, Nivan Nelson, Abhiram Karukayil, Rohith Kulothungan, Mohammed Afil Farhan, Birthe Nesset, Meriam Moujahid, Tanvi Dinkar, et al. 2023. Furchat: An embodied conversational agent using llms, combining open and closed-domain dialogue with facial expressions. *arXiv preprint arXiv:2308.15214*.

Charles R Crowelly, Michael Villanoy, Matthias Scheutzz, and Paul Schermerhornz. 2009. Gendered voice and robot entities: perceptions and reactions of male and female subjects. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3735–3741. IEEE.

Nicole Dodd, Michelle Cohn, and Georgia Zellou. 2023. Comparing alignment toward american, british, and indian english text-to-speech (tts) voices: Influence of social attitudes and talker guise. *Frontiers in Computer Science*, 5:1204211.

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

Leon Hanschmann, Ulrich Gnewuch, and Alexander Maedche. 2024. Saleshat: A llm-based social robot for human-like sales conversations. In *Chatbot Research and Design*, pages 61–76, Cham. Springer Nature Switzerland.

Stephen M Kromka and Alan K Goodboy. 2019. Classroom storytelling: Using instructor narratives to increase student recall, affect, and attention. *Communication Education*, 68(1):20–43.

Julia Simkus. 2023. Within-Subjects Design: Examples, Pros & Cons. https://www.simplypsychology.org/within-subjects-design.html. [Accessed 22-04-2024].

Philipp Ulsamer, Kevin Pfeffel, and Nicholas H Müller. 2019. Indoor navigation through storytelling in virtual reality. In *Learning and Collaboration Technologies. Ubiquitous and Virtual Environments for Learning and Collaboration: 6th International Conference, LCT 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21*, pages 230–239. Springer.

Michael BW Wolfe and Joseph A Mienko. 2007. Learning and memory of factual content from narrative and expository text. *British Journal of Educational Psychology*, 77(3):541–564.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Dong Yu and Lin Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. 2023. Large language models for human–robot interaction: A review. *Biomimetic Intelligence and Robotics*, 3(4):100131.

Zhonghan Zhao, Wenhao Chai, Xuan Wang, Li Boyi, Shengyu Hao, Shidong Cao, Tian Ye, Jenq-Neng Hwang, and Gaoang Wang. 2023. See and think: Embodied agent in virtual environment. *arXiv preprint arXiv:2311.15209*.