



# UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering,  
Mathematics & Computer Science

## Gender Perceptions in chatbots: The Role of Linguistic features

Meenakshi Prakash  
M.Sc. Thesis  
September 2024

---

**Supervisors:**

dr. K.P. Truong  
dr. M. Theune  
dr. J. Karreman  
H. Garcia Goo MA

University of Twente  
P.O. Box 217  
7500 AE Enschede  
The Netherlands

---

# Abstract

Many conversational agents are assigned female gender, which, coupled with their roles as assistants and the language they use, can reinforce gender stereotypes. Exploring and designing agents and those that challenge these stereotypes is crucial for ethical design that focuses on gender equality. This study investigates the factors that contribute to the perception of gender in conversational agents, focusing specifically on linguistic features. While designers often gender chatbots through names, pronouns, and avatars, research has shown that less explicit linguistic cues within the conversation can also influence gender perception. Despite extensive research on gender-based linguistic differences, the impact of specific linguistic features on chatbot gender perception has not been thoroughly examined. This study addresses this gap by examining how linguistic features like intensifiers and emotional language, which have been observed at higher rates in female speech and writing, influence gender perception. Using a between-subjects design, the experiment involves three chatbots: one incorporating intensifiers, one incorporating emotional language, and a control chatbot without these features. The results show a slight trend in the expected direction, suggesting that linguistic features might influence participants' gender perception of the chatbot. However, these findings are not statistically significant. The trend points to the potential value of further study to explore whether linguistic features affect gender perception. Given the subtle nuances of language, more research is necessary to gain a deeper understanding.

# Acknowledgement

This thesis has been a LEGO set built by a group of people. It took almost eight months, the better part of a year, and I have many to thank.

First, my parents and my grandmother. They supported me through thick and thin from across oceans and stood by me even when I was not sure about myself. Thank you for tolerating me during those times when I am sure it would not have been pretty to watch. And thank you for your prayers.

My aunt and uncle for constantly giving me the motivation to continue and cheering me up during our occasional calls.

My friend, Hridya. You were in the same condition as me and helped me pick myself up and move forward. You were one of the few people I could talk to and be sure to be understood amidst this. I could not have done it without your help and I cannot thank you enough.

My supervisors, Khiet Truong, Mariet Theune, and Hideki Garcia Goo, for being patient with me always. They guided me, listened to my ideas, and helped me as much as they could in completing this journey. I loved our discussions, which opened my horizons on inclusivity and design. Thank you, Joyce Karreman, for joining our team at the last minute. I hope you enjoyed this endeavor, and I highly appreciate your involvement in my thesis.

Last but not least, I want to thank the universe for helping me survive through this. I derived joy and energy from you to move forward.

# Contents

- Abstract** **ii**
- Acknowledgement** **iii**
- 1 Introduction** **1**
- 2 Literature Review** **4**
  - 2.1 Gender in Chatbots . . . . . 4
  - 2.2 Communication styles seen in men and women . . . . . 5
- 3 Chatbot Development** **13**
  - 3.1 Creating Gender-Neutral Role . . . . . 13
    - 3.1.1 Pre-study to confirm the gender neutrality of the chatbot role. . 14
  - 3.2 Conversation design . . . . . 15
    - 3.2.1 Conversation objectives . . . . . 15
    - 3.2.2 Conversation Structure . . . . . 15
    - 3.2.3 Dialog Flow and Conversation Management . . . . . 17
  - 3.3 Integration of Linguistic Features . . . . . 20
    - 3.3.1 Control Chatbot . . . . . 20
    - 3.3.2 Intensifier Chatbot . . . . . 20
    - 3.3.3 Emotional Chatbot . . . . . 21
- 4 Research Methodology** **25**
  - 4.1 Experimental Design . . . . . 25
  - 4.2 Participant Recruitment . . . . . 26
  - 4.3 Survey design and Measurements . . . . . 27
    - 4.3.1 Demographics . . . . . 27
    - 4.3.2 Experience with Chatbots . . . . . 27
    - 4.3.3 Gender Perception . . . . . 28
    - 4.3.4 Assessing the manipulations . . . . . 29
    - 4.3.5 Anthropomorphism . . . . . 29
  - 4.4 Procedure . . . . . 29

---

4.5	Data analysis . . . . .	30
<b>5</b>	<b>Results</b>	<b>34</b>
5.1	Perceptions of Chatbot Gender . . . . .	34
5.1.1	Perceived masculinity, femininity and gender neutrality . . . . .	34
5.1.2	Short form Bem Sex Role Inventory . . . . .	36
5.2	Anthropomorphism . . . . .	38
5.2.1	Perceived Naturalness . . . . .	38
5.2.2	Perceived Humanlikeness . . . . .	39
5.3	Relationship between Gender of the participants and Gender percep- tion across all chatbots . . . . .	40
5.4	Results that deal with multiple measures-Gender perception and An- thropomorphism . . . . .	41
5.4.1	Name assignment . . . . .	41
5.4.2	Participant's Description of the Chatbot . . . . .	42
5.5	Assessment of Validity of Manipulation . . . . .	45
5.5.1	Feature Detection in Chatbots . . . . .	45
5.5.2	Participant Perceptions of Chatbot Attributes . . . . .	46
<b>6</b>	<b>Discussion and Conclusion</b>	<b>49</b>
6.1	Discussion . . . . .	49
6.1.1	Gender Perception . . . . .	49
6.1.2	Anthropomorphism . . . . .	51
6.2	Chatbot design implications . . . . .	51
6.2.1	Feature integration . . . . .	51
6.2.2	Conversation design . . . . .	52
6.3	Limitations . . . . .	53
6.4	Contribution . . . . .	53
6.5	Future Research Directions . . . . .	54
6.6	Conclusion . . . . .	54
	<b>References</b>	<b>56</b>
	<b>Appendices</b>	
<b>A</b>	<b>Gender-Specific Linguistic Features</b>	<b>64</b>
<b>B</b>	<b>Dialogues</b>	<b>70</b>
B.1	Example Dialogues . . . . .	70
B.1.1	Control Chatbot . . . . .	70
B.1.2	Intensifier chatbot . . . . .	71

---

B.1.3 Emotional chatbot . . . . .	73
<b>C Experiment materials</b>	<b>75</b>
C.1 Information sheet and Consent form . . . . .	75
C.2 Debriefing . . . . .	76
<b>D Participant Demographics</b>	<b>77</b>
<b>E Survey results</b>	<b>78</b>
E.1 Names assigned by participants for the chatbots . . . . .	78
E.2 Participants' description of chatbots . . . . .	79
E.2.1 Control Chatbot . . . . .	79
E.2.2 Intensifier Chatbot . . . . .	81
E.2.3 Emotional Chatb . . . . .	82

## Introduction

Conversational agents (CA) both text and voice-based are becoming more common in our society. They have been utilized for several purposes from education [1], healthcare [2] [3], customer service [4] to companionship [5], thus playing multiple roles in the current society. While interacting with these agents we subconsciously attribute human attributes to them. The Computers Are Social Actors framework (CASA) often used in human-computer interaction studies states that the relationship between humans and computers is social and humans interact with computers using the same rules and behaviors they use for social interactions [6] [7]. These attitudes and behaviors can lead to the perception that machines have human-like qualities, such as personalities [8].

This perception could be influenced by the design of the agents. A UNESCO report from 2019 reveals that most voice-based CAs are designed to be female exclusively or female by default [9]. A recent study [10] looked into 1375 chatbots from chatbots.org in the fields of branded conversation, sales, and customer service. It concluded that gender bias is particularly evident in these fields as 77% of chatbots were classified as female by default or an only option. Another study that looked into over 300 chatbots, assistants, and AI movie characters reveals that 35% of chatbots and 67% of assistants are female [11]. West et al. [9] concluded in the above-mentioned UNESCO report that the prominence of female-sounding voice assistants encourages stereotypes of women as submissive and compliant.

There have also been studies examining the benefits of assigning a gender to conversational agents. A study from 2020 [12] found that female chatbots are preferred over male chatbots because they are perceived as more human and more likely to consider the unique needs of the user which leads to a favorable attitude toward AI solutions. Women are ascribed qualities of warmth which also apply to female-gendered robots [12]. This ascription reflects traditional views of women as inherently warm [13]. There are two sides to this coin. On one hand, assigning female characteristics to chatbots can enhance the product's usability. On the other

hand, it risks reinforcing harmful societal stereotypes.

Additionally, the perceived benefits seen above associated with gendered design often hinge on reinforcing these stereotypes. This prompts us to address the question of whether we truly need to gender conversational agents. Studies have shown that gender does not significantly affect trust or competence ratings of AI assistants [14]. In the case of social robots perceived occupational competency was found to be a better predictor for human trust rather than robot gender [15]. Therefore it is valuable to look into creating non-gendered agents and those that subvert the gender stereotypes, for which we need to understand those stereotypes and how gender manifests in agents.

Gender specification in conversational agents is done through different methods. For voice-based agents, voice, name, an avatar (which is a virtual character generated to personify the entity), and descriptions using gendered pronouns can be major factors. For chatbots too the most common ways are names, avatars, and descriptions using gendered pronouns. A major part of a chatbot is its conversational style. Language and gender have been researched for decades and there have been several findings on differences in language related to gender. A study by Vanderlyn et al. [16] which looked into how the linguistic style of an agent affected the likability and perception of the agent, found that linguistic style alone influences user perception of the chatbot's gender. When an empathetic style was used, the participants projected their own gender to the chatbot which helped to mitigate negative stereotypes associated with one gender over the other. This was also viewed as an alternative to explicitly gender-neutral chatbots.

However, this area of research needs further exploration to identify which specific features influence the perception of gender. Research has been done on linguistic differences between languages used by men and women. Various features have been found to distinguish the language. The use of intensifiers has been a prominent feature in the female language [17] [18] [19] [20] [21]. Similarly works have also found higher amounts of words describing emotions, social, psychological processes, and family in female language [22] [20] [23] [24]. These features could also translate to chatbots and influence the perceived gender of chatbots.

This research investigates the potential influence of these linguistic features on the perceived gender of a chatbot, considering factors beyond the more direct indicators such as name or avatar. The research question guiding this investigation is: **How does the use of linguistic features, specifically intensifiers and emotion-inducing words, influence the perception of gender in text-based conversational agents?**

It is interesting to understand which of these features has a stronger impact on



the gender perception. This will be answered by the research question:

**How does the effect of intensifiers on the perceived gender of text-based conversational agents compare to the effect of emotion-inducing words?**

As these features were found in higher amounts in language used by women, the following hypotheses were constructed.

**H1: The usage of intensifiers by the chatbot results in increased perception of the chatbot as feminine**

**H2: The usage of emotion-inducing words by the chatbot results in increased perception of the chatbot as feminine**

To address these questions and test the hypotheses, we performed an experiment where the participants interacted with chatbots that were incorporated with intensifiers and emotion words.

Context and roles also influence the perception of gender. Stereotypes related to jobs and gender often lead to assumptions that certain jobs are suited for a particular gender. To prevent the role from affecting gender perception, a gender-neutral role is chosen for developing the chatbot. Research indicates that the role of a travel guide has minimal gender associations [15]. This will also be addressed in subsequent chapters of this thesis.

The structure of this thesis is as follows: Chapter 2 reviews the existing literature about gender in chatbots and the relationship between gender and language. Chapter 3 details the development of the chatbot, covering the technical aspects and design decisions made during its creation. Chapter 4 outlines the methodology used in this research, including the experiment design and methods for data collection and analysis. Chapter 5 presents the results of the study, providing an analysis of the data collected. Finally, Chapter 6 discusses the findings, concluding with their implications and suggestions for future research.

# Literature Review

## 2.1 Gender in Chatbots

Chatbots have been developed for various purposes. Voice-based conversational agents like Alexa, Siri, and Google Assistant serve as personal assistants, while text-based ones are used in applications such as company websites, messaging applications, personal assistants, companionship, and even therapy.

Chatbots are designed to emulate human behavior in terms of visual and conversational cues [25]. Most chatbots represent female characteristics as default settings. This is expressed mostly through names, avatars and gendered pronouns. When chatbot avatars are gendered, people expect them to possess gendered knowledge and characteristics, associating stereotypes to them: warmth, nurturing, and helpfulness for women, and competence, authority, and agency for men.

Bastiansen et al. [26] performed a study to understand how warmth in language and a chatbot's assigned gender elicit stereotypes and if that could affect the perceived trust, competence, and helpfulness users assign to the chatbot. They hypothesized that a female chatbot using warm language would be perceived as more trustworthy and helpful but less competent. They also aimed to investigate whether a female chatbot would be penalized for having features not stereotypically associated with the gender. They gendered the chatbots using gendered names, pronouns, and emojis. The results did not show significant effects of warmth on competence, trust, or helpfulness. Nor did they show any interaction effect between warmth and gender. They suggested several reasons for these findings: implicit stereotypes that people might try to correct, warmth not being strong enough, short conversations, participants not seeing the chatbot as human, and context dependency.

Gender cues like voice can also affect the interaction experience. Guo et al. studied the effect of chatbot voice gender on customers' intention to pay debt. They modulated the voice to incorporate stereotypical features: forceful and assertive for male agents, gentle and compassionate for female agents. The results demon-

strated that masculine attributes had significant negative effects on both male and female customers, while feminine attributes had significant positive effects only on male customers [27]. Another study examined the gender affordances of conversational agents using nine chatbots (six embodied, two disembodied, and one robot). Gender was indicated through images and names of the chatbots. The results showed that female-presenting chatbots faced more negative stereotypes, sexual attention, and verbal abuse compared to male-presenting ones [28]. Thus design involving the manipulation of gender can significantly impact user experience, influencing perception, but it may also occasionally reinforce stereotypes.

When creating gender-neutral chatbots or ones that reduce stereotyping, various factors should be considered. Avoiding explicit gender markers might not be the only method. A recent study examined the impact of linguistic style on the likability and perceived gender of chatbots. They did not use any gender markers. However, it was found that a passive style led participants to perceive the chatbot as male, while an empathetic style was perceived by 35% of participants as the same gender as themselves. The empathetic style ranked high in likability, whereas a direct and assertive style elicited gender stereotypes among participants. The study concluded that using an empathetic, non-gender-expressive style could reduce stereotypes in chatbot interactions [16]. In the case of gender-neutral voice research, Yeon et al. investigated whether a gender-neutral AI as a speech-based conversational agent would be preferred. Participants did not favor this option because the agent's identity was challenging to grasp [29].

The findings across the studies discussed above are varied. In some cases, gender has no effect on perceived trust, competence, and other attributes, while in other cases, gender specification improves user interactions. Female chatbots often receive negative treatment and stereotypes. However, when gender-neutral voices are introduced, they can make the interaction less user-friendly due to user confusion and a less human-like agent. Most studies that explore how gender impacts chatbot conversations specify gender through visual cues, names, gendered pronouns, and sometimes linguistic features like warmth and empathy. Despite ample studies on conversational differences between men and women, this has not been extensively examined in chatbots.

## **2.2 Communication styles seen in men and women**

When examining gendered chatbots, distinctions are often made through roles, names, and avatars. However, their communication styles and linguistic features may also contribute to perceived gender differences. This phenomenon can be better understood by exploring the linguistic variations in the text and speech pat-

terns of different genders. Numerous studies have investigated language differences that are seen with gender, with some arguing that there are no significant disparities [30] [31], while others have demonstrated variations in language use [32]. These differences may manifest in terms of several linguistic features like lexical choice and syntactic patterns. However, it is imperative to approach these findings within societal frameworks rather than individual levels. That is, the findings may not generalize to all individuals of a particular gender. Furthermore, it is essential to acknowledge that much of this research still operates within a binary gender framework, despite the evolving view of gender as a spectrum rather than a binary construct [33].

Various methodologies have been employed to analyze language differences between men and women across different contexts ranging from professional and task-oriented to casual conversations and stream of consciousness, where participants express their thoughts and feelings freely without additional guidance. The outcomes of these studies have been diverse, with varying conclusions emerging. Here we are considering only studies of the English language.

The linguistic features found in gender-linked language research are of different types and can be categorized into semantic, syntactic, pragmatic, speech acts, and lexical. A summary of the features and the studies that examined them is presented in the table A.1 in Appendix A.

### **Syntactic features**

Syntactic features focus on the structure and arrangement of words within sentences. Studies have found that several syntactic features, such as those listed below, occur more frequently in the language of one gender compared to the other. These findings span various contexts and age ranges. A few prominent syntactic features found in gendered language are listed below.

- **Progressive verb form** - The progressive verb form consists of the "-ing" form of the verb, such as in the phrase "I am writing." This linguistic feature has been observed to occur more frequently in men during public speaking scenarios [20]. It appears more often in women when describing photographs [22] and in high school impromptu essays [34]. Overall, this linguistic feature has been found to be higher in female language in two studies and in male language in one study.
- **Mean Sentence Length** - The mean sentence length, calculated by dividing the number of words by the number of sentences, was higher in the language used by women across various contexts such as essays [35] [34], descriptions of photographs [22], public speaking [20], oral descriptions of photographs [19],

interviews [36], and during roleplay scenarios involving managerial criticisms directed towards employees [21]. However, a higher mean sentence length was also observed in male kids' writing in fourth-grade essays [34]. Overall, this linguistic feature has been found to be higher in female language in six studies and in male language in one study.

- Sentence initial adverbials, which modify verbs, adjectives, or other adverbs to offer extra details about time, manner, place, or frequency, are positioned at the beginning of a sentence. They answer the question of when, where, how, or to what extent the action described in the main clause occurs. For instance, "*Instead of being the light blue, it is.*" This syntactic characteristic was found in a higher frequency in female writing across various contexts, including written descriptions of photographs [22], fourth-grade essays [34], and speech during public speeches [20], dyadic interactions [17], and roleplays of managerial criticisms [21]. Overall, this linguistic feature has been found to be higher in female language in five studies.
- Elliptical sentences - Elliptical sentences are units in which either the subject or predicate is understood and thus omitted, typically starting with a capital letter and concluding with a period. For instance, "*Gorgeous!*" or "*A beautiful snowy setting.*" This syntactic feature was observed in a higher frequency in male writing within the context of written descriptions of photographs [22] and in speech during oral descriptions of the same [19]. Overall, this linguistic feature has been found to be higher in male language in two studies.
- Dependent clauses- Dependent clauses are clauses that serve to specify or qualify the words that convey primary meaning. For example, "*which is mostly covered,*" "*where the shadows are,*" "*in which something*". Dependent clauses were observed at a higher frequency in female writing across essays [34] [35], descriptions of photographs [22], speech during interviews [36] and roleplay scenarios involving the giving of criticisms by managers [21]. Overall, this linguistic feature has been found to be higher in female language in five studies.
- Common nouns: Common nouns are generic names for persons, places, things, or ideas that belong to a class or group. For example, "*The car was in the parking lot.*" "*The dog was mine.*". These are observed more frequently in male writing indicating that male authors were likely to specify the things that they write about or that male writers mention classes of subjects that they are writing about. This was in contrast to female writers who personalize and use pronouns to link the subjects that they write about For example, "*It was in the parking lot.*". [37].

- Noun specifiers - Noun specifiers are words that provide more specific information about a noun. They can include articles (such as "a," "an," and "the"), demonstratives (such as "this," "that," "these," and "those"), possessives (such as "my," "your," "his," "her," "its," "our," and "their"), and quantifiers (such as "some," "many," "few," "several," "all," and "none"). Noun specifiers are used by authors to provide information on things that they assume readers do not know and thus are a feature of "informative" writing observed more frequently in male authors' writing [37].
- Oppositions- Oppositions are a linguistic feature where a speaker or writer initially presents a statement and then contrasts it with another statement that has an opposite or differing meaning. For example, "*The snow must have fallen fairly recently, but it has been a while*" "*very peaceful, yet full of movement*". This phenomenon is observed more frequently in female speech during oral descriptions of photographs [19], public speeches [20] and in male speech during roleplay scenarios involving criticisms by managers [21]. Overall, this linguistic feature has been found to be higher in female language in two studies and in male language in one study.
- Intensifiers - Intensive adverbs or intensifiers are adverbs that modify the degree or intensity of the action or quality being described. This linguistic feature is typically utilized to convey the meaning of the modified adjective or verb more strongly and has been identified in empirical studies as a characteristic more frequently observed in female writing. Examples include "*really*," "*so*," "*very*," and "*quite*." Intensifiers has been observed more frequently in female speech during dyadic interactions [17] [18] [38], group discussions [39], oral descriptions of photographs [19], public speeches [20], and roleplay scenarios involving the criticisms by managers [21]. It was also observed frequently in women's writing in an extensive analysis of 14,000 text samples from 70 studies [40] and in social media messages on Facebook [41]. Overall, this linguistic feature has been found to be higher in female language in nine studies.
- Negations - Negations involve expressing the opposite of a statement and are achieved using words like "*not*," "*no*," and "*never*." It consists of stating what something is not present or happening. For example, ("You don't feel like looking"). This feature is observed frequently in female speech compared to male speech during oral descriptions of photographs [19] and public speeches [20] and in male speech during roleplay scenarios involving criticisms by managers [21]. Overall, this linguistic feature has been found to be higher in female language in two studies and in male language in one study.

### **Semantic features**

Semantic features are theoretical units of meaning-holding components used to represent word meaning. They focus on phrases, sentences, overall discourse [42] and into the conceptual context of these entities. Several semantic features have been observed as identifiable features of gendered language.

- Reference to quantity or Quantifiers - Quantifiers involve phrases that reference an amount or quantity, such as “*below 320 F,*” “*most of the area,*” and “*6–8 thousand feet elevation.*” This linguistic feature has been observed more frequently in male speech during event descriptions, oral descriptions of photographs [19], interviews, and oral descriptions of pictures. It has also been noted in male writing in event description essays [22] [23]. Overall, this linguistic feature has been found to be higher in male language in five studies.
- Locatives - Locatives are phrases typically indicating the location or position of objects, such as “Right next to the” or “in the background.” This linguistic feature has been observed more frequently in male speech during oral descriptions of photographs [19] and event descriptions [22]. Overall, this linguistic feature has been found to be higher in male language in two studies.
- Reference to emotion involves any mention of an emotion or feeling, such as “happy,” “enticing,” or “depressing.” This linguistic feature is observed more frequently in and was an indicative feature of female writing [40], written descriptions of photographs [22], social media messages [24]. It was also observed frequently in female speech during public speeches [20] and descriptions of pictures [22]. However words expressing negative emotions was observed higher in male writing in a study examining natural conversations of students [43] and positive emotions in female conversations [41]. Emotion words were also observed more frequently in male speech during managerial criticisms which is contradictory to the previous research [21]. Overall, this linguistic feature has been found to be higher in female language in eight studies and in male language in two studies.

### **Lexical features**

Lexical features consist of word length, word frequency, and high-frequency words [44]. Language used by different genders is seen to have patterns in terms of the type of words used. Although several words are indicative of gender in language, they are often found in groups that convey specific meanings. Therefore, most of them are categorized as semantic features.

- **Pronouns:** Pronoun usage has been identified in both genders. First-person singular, second-person, and third-person pronouns are recognized as indicative features of female language [37]. Personal pronouns (I, you, we) were observed in higher numbers in female language, particularly in speech contexts such as event descriptions, recounting tales and free play, oral descriptions of photographs [19], dyadic interactions [17], interviews [36], and written contexts like stories and newspaper articles [40]. The use of "I" in the subjective case was also noted in male writing, specifically in descriptions of photographs [22] and fourth-grade impromptu essays [34]. Overall, this linguistic feature has been found to be higher in female language in five studies and in male language in two studies.

### **Pragmatic features**

Pragmatics focus on the implied meaning of a feature. Certain pragmatic features appear more frequently in the language used by women and men in specific contexts. These are discussed below.

- **Assertive vs Affiliate language:** Assertive language includes directive statements, giving information and disagreeing with or criticizing other's contributions. This serves as a way to advance one's personal agency in a given situation. Affiliate language involves showing support, expressing agreement and acknowledging the partner or other person's contribution. This communication style functions to affirm or positively engage with the other person. Reviewing multiple studies focusing on the same reveals that assertive language was seen more in male speech and affiliate in female [45].
- **Hedges -** Hedges, expressions used to soften the impact of a statement or make it less direct, might indicate a lack of confidence or diminished assuredness. They include modifiers like "kind of," "a little," and "somewhat." Lakeoff identifies this linguistic feature as indicative of female language. It is observed more frequently in female speech during dyadic interactions [18] and in writing, particularly in fourth-grade impromptu essays [34] [23]. Overall, this linguistic feature has been found to be higher in female language in five studies.
- **Uncertainty verbs -** Uncertainty verbs are verb phrases that indicate a lack of certainty or assuredness, such as "I wonder if," "Seems to be," or "I'm not sure" This linguistic feature has been observed frequently in female speech during interviews [36] and in writing in descriptions of photographs [22]. [23] and speech [46]. Overall, this linguistic feature has been found to be higher in female language in three studies.



- Tag questions - Tag questions are phrases that are added to the end of a statement turning it into a question, generally used to seek confirmation or agreement or sometimes indicate uncertainty or politeness. For example "You're coming to the party, *aren't you?*". Tag questions were claimed to be a linguistic feature indicative of female language by Lakoff, There has been evidence for and against this claim. In the female speech, tag questions were found more frequently in problem-solving groups, dyadic interactions [18], oral descriptions of photographs [19], and interviews. However in male speech, this feature was also observed more frequently under the context of conference participation and informal conversations [23]. Overall, this linguistic feature has been found to be higher in female language in three studies and in male language in one study.
- Fillers - Fillers or words or sounds used to fill gaps in a conversation without apparent semantic intent, such as "*like*" and "*you know*", have been claimed as a linguistic feature indicative of female speakers by Lakoff. However, there is contradicting evidence from two studies against and three in favor. In studies, fillers were observed prominently in various contexts, including female speech during public speeches [20] and writing in fourth-grade essays [34], as well as male speech during dyadic interactions [17] and oral picture descriptions [19] [23]. Overall, this linguistic feature has been found to be higher in both male and female language in two studies.
- Extra polite forms- Lakoff claimed that women were more likely than men in the same situation to use extra-polite forms (e.g., "Would you mind ... "), a claim that was supported by subsequent empirical work [39] which tested this hypothesis in a same-sex and mixed group problem-solving activity.

### Speech acts

Speech acts are utterances serving specific purposes in communication. Some examples of this are questions, apologies, greeting, complaints, and invitations.

- Questions - Questions are expressions used to request information, response, or clarification from someone. This feature was [17] found more commonly in women's contributions to dyadic interactions [17] and couples conversations. (e.g., "*Does anyone want to get some food?*"). Overall, this linguistic feature has been found to be higher in female language in two studies and in male language in one study. This feature does not include directives in question form.

- Directives - Directives are sentences or phrases that tell the partner in conversation to do something (e.g., "Let's go get some food"). These were more likely found in men's speech during dyadic interactions [17] and interviews [47] and in female speech during managerial criticisms [21]. Overall, this linguistic feature has been found to be higher in male language in two studies and in female language in one study.

In summary, various linguistic features have been identified as distinguishing elements of the language based on gender. Consequently, these features could influence gender perception when incorporated into chatbots. Intensifiers were found to be an indicative feature and observed more frequently in female language in nine studies. Emotion words expressing emotion and social processes were also an indicative feature and found more often in female language in eight studies. The usage of intensifiers and emotional words emerged as particularly prominent. Therefore, the next step involves integrating these linguistic characteristics into chatbots to evaluate their effect on gender perception.

# Chatbot Development

The chatbots for the experiment were developed using Google Dialogflow, which offers a user-friendly interface and an inbuilt language model [48]. Several decisions were made during the design of the chatbots.

## 3.1 Creating Gender-Neutral Role

The role of the chatbot influences gender perception [49] [15]. While creating the chatbot, a critical part of the design included which field the chatbot should operate in and which role it should assume. To understand the effects of the linguistic features, a gender-neutral role that did not impact the participant's perception of the chatbot's gender was utilized. To do this we looked into what stereotypes prevailed in social roles.

Traditional beliefs often assert that due to stereotypical traits and temperaments, men and women are suited for different occupations and roles. In a 1975 empirical study, students were found to associate certain occupations with specific masculine or feminine traits [49]. There have been changes in these stereotypes as more occupations have become gender-neutral. This is true for those roles where the male-to-female ratio has become less and thus more balanced. However, the occupations where this is not the case still have gender stereotypes attached to them which can lead to discrimination and dire societal issues if perpetuated.

People often apply gender stereotypes to robots that exhibit stereotyped gender characteristics [50]. Research has shown that when participants viewed an image of a robot with a specific haircut, they attributed traditional gender traits to it. For instance, a robot with a female haircut was seen as compassionate and empathic, making it seem more suited to roles such as patient or child care. Conversely, a robot with a male haircut was perceived as assertive and competitive, aligning it with tasks like repairing equipment and transporting goods [50]. In the case of chatbots with no

avatars, these physical features are not applicable. However, it can be understood that there are stereotypes of child and patient care as feminine roles, and repairing and transporting goods as masculine ones.

In a 2020 study investigating the impact of occupational stereotypes on trust towards social robots, 50 participants from the United States, aged 22 to 72, were tasked with categorizing 14 occupations commonly associated with social robots according to the gender they typically associate with each role. Participants were given the option to select neither gender. Occupations perceived as predominantly female included nanny (98%), receptionist (92%), nurse (94%), home health aide (94%), teacher (78%), and therapist (64%). Occupations predominantly associated with males included firefighter (94%), security guard (94%), package delivery person (84%), and surgeon (70%). Those perceived as neither gender included tour guide (48%) and news anchor (42%) [15]. Tour guide is one of the roles that emerges with the least stereotypes attached to it. Apart from this travel and tourism is one of the sectors which maintain a fairly balanced employment ratio (54% women to men) between men and women according to the WTTC report [51]. This is also a promising field for chatbot applications. Discussing travel is relevant and engaging for many people, making it an ideal topic for chatbot interactions. Therefore this field was chosen to develop the chatbot.

### **3.1.1 Pre-study to confirm the gender neutrality of the chatbot role.**

To confirm if the field indeed was gender-neutral, 10 people were asked what gender came to mind when discussing travel experiences. This question was posed generally, without reference to any specific medium, such as chatbots. The participants were university students from diverse cultural backgrounds, with seven being women, recruited through convenience sampling. Careful consideration was taken to not associate any terms like assistant, advisor etc to infuse any gender perceptions to the role. The pre-study was informal and orally conducted. The participants were recruited through convenience sampling. Out of 10 participants, 8 people said they don't necessarily see or picture a gender when they think of the scenario. 2 people said they thought of their own gender when imagining the scenario. This led to the finalization of the field of the chatbot and the further conversation design.

## 3.2 Conversation design

### 3.2.1 Conversation objectives

Designing the conversation involved several key decisions to ensure it fulfilled the research objectives- to effectively portray linguistic features to understand the effect of those on gender perception.

One of the key decisions was to keep the conversation informal. Mulac et al. proposed that informal conversations are more likely to exhibit differences in linguistic features among genders [22].

Determining the appropriate conversation length was an important factor in designing the interaction. A study examining participants' ability to perceive the personality of their conversation partner found that a 5-minute conversation was sufficient [52]. This finding guided our decision on the duration of the chatbot interactions. Another major consideration was to let all the users have a uniform exposure to the manipulations.

Initially, a conversation was designed to help users find their ideal travel destination. However, it became clear that the length of such a conversation would vary depending on user responses, leading to inconsistent exposure to the manipulations among participants and varying experiences with the linguistic style of the chatbot. Additionally, the close-ended nature of the questions made the conversation monotonous and formal. Planning a trip might also remind people of the role of a personal assistant, which is often linked to gender.

Therefore the final aim of the dialogue and interaction design was to have a casual conversation about travel with no particular goal other than discussing travel experiences. This approach ensured that the conversation length remained consistent, regardless of the participant's responses or decisions during the interaction which was one of the main goals.

### 3.2.2 Conversation Structure

The current design incorporates a linear conversation flow. This means that the conversation progressed through a defined sequence of steps regardless of the user responses. This is to maintain fairly consistent exposure across participants regardless of their responses as expressed before. An example of a part of the dialog flow is provided in figure 3.1.

The chatbot uses open-ended questions such as "What are your thoughts on traveling?" "Moving on, In new locations language barriers can be quite challenging. How do you manage them?" and "How do you typically select your destinations?". These questions are meant to encourage a casual, informal discussion about travel.

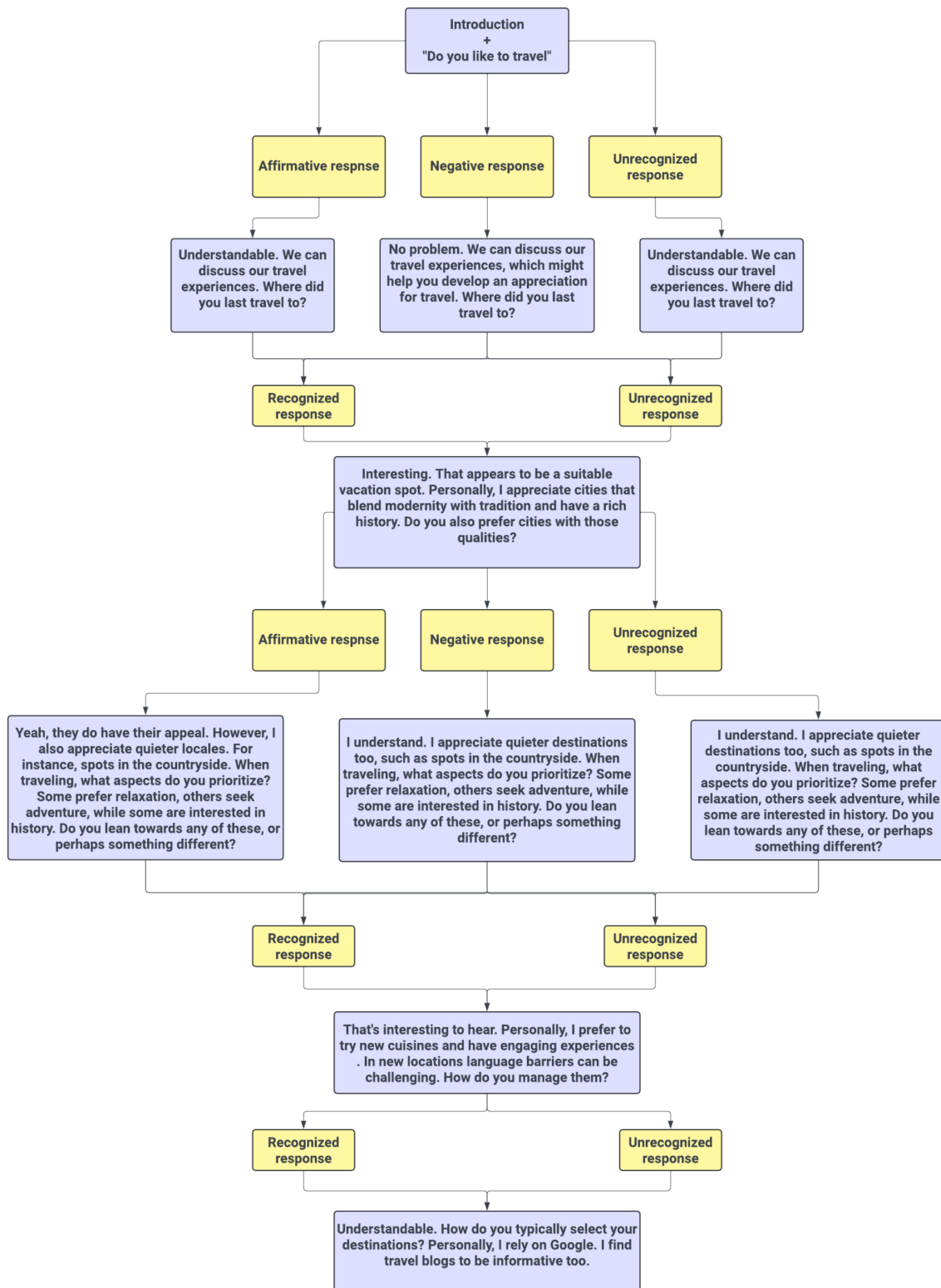


Figure 3.1: Dialogue flow

There are 10 such questions throughout the dialogue of which the conversation flows up to question 5 is provided in figure 3.1. The user's answers to these questions are not responsible for any major changes in the dialog flow as seen in the diagram.

For instance, while the user's response might slightly influence the acknowledgment part of the reply, it does not alter the overall flow or outcome of the conversation. The only part of the chatbot dialogue that is different according to the user response is the initial acknowledgment phrases, like "Understandable" or "I get it" or "I feel you" for an affirmative response and "Not a problem" or "I get it. However.." in case of negative responses. This is identified with the help of intents which will be explained in the subsection below. Unknown responses are designed to be responses to have neutral acknowledgment phrases that neither indicate agreeing nor disagreeing but just acknowledging. This design allows the chatbot to continue the linear path of discussion while still making the user feel like their input is being acknowledged.

The chatbot's responses to the user's answers are designed to be deliberately generic and versatile. Phrases like "Interesting," "I understand," "Understandable," "Totally get it," "That is really interesting to hear," and "I see where you are coming from" are employed to acknowledge a range of user inputs. This approach allows the chatbot to effectively manage open-ended questions, even when the user's responses are unpredictable.

### 3.2.3 Dialog Flow and Conversation Management

The development of the chatbot began with defining intents, which are specific objectives or goals that the user seeks to achieve through their input. By recognizing these intents, the chatbot can interpret user input and provide appropriate responses. In the current implementation, each intent represents a unique stage in the conversational flow. A total of 11 dialogue exchanges have been established between the chatbot and the user, with each exchange being associated with a distinct intent or two depending on the type of question. For example, in response to the question "Do you like to travel?" the user's reply can be positive, negative, or unrecognized. The chatbot identifies the intent through training phrases. Upon recognizing a particular intent, the chatbot responds with the corresponding response tailored to that intent. In the third chatbot response (Figure 3.1), both recognized and unrecognized inputs lead to the same response, even though they are treated as different intents. This setup is due to the way Dialogflow operates, where it distinguishes between recognized and fallback intents. Even if the output is the same, Dialogflow still requires separate intents to manage recognized inputs and provide fallback responses when the input is not recognized, ensuring smooth conversation

flow regardless of input understanding.

A key feature of the chatbot's functionality is its ability to maintain context throughout the conversation. In this study, context which is a feature of Dialogflow is used to control the flow of the dialogue, ensuring that the conversation follows the intended sequence from start to finish. Essentially, context acts like a marker or code that is generated whenever an intent is recognized.

When the chatbot recognizes the first intent, it triggers an output context linked to the first question. This output context then becomes the input context for the next step in the conversation, ensuring that only the correct next intent is recognized. For example, the second intent will only be recognized if the output context from the first question is active. This means that each subsequent intent is dependent on the context generated by the previous step, allowing the chatbot to guide the conversation precisely as needed for the research.

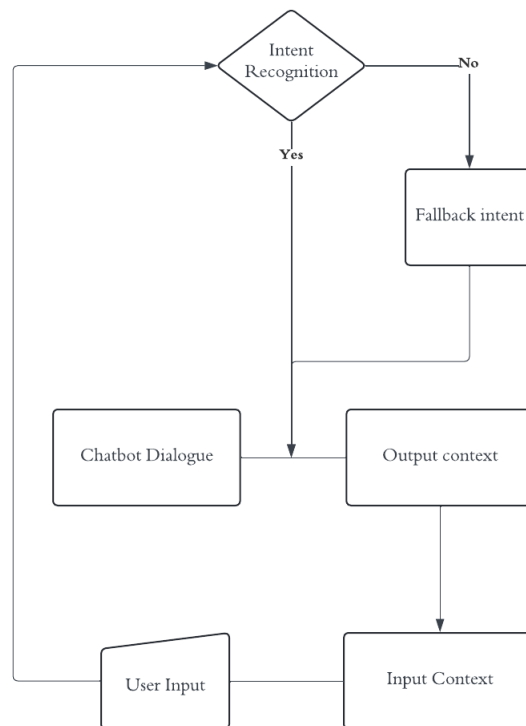
A significant feature of this design is the use of fallback intents, another feature of Dialogflow that handles unrecognized inputs. Fallback intents were used whenever a user's response didn't match any of the predefined intents. Since the main goal was to keep the conversation on its fixed path, fallback intents were designed to redirect the dialogue back to the intended flow when no specific intent was recognized. The chatbot's acknowledgments in these fallback cases were deliberately generic, ensuring that they did not contradict the user's input while smoothly continuing the conversation. The figure 3.2 illustrates how intents are recognized, how contexts are managed to maintain the flow, and how fallback mechanisms redirect the conversation when necessary.

To summarise, when a user provides an input, the chatbot checks if it matches the current context and intent. If it does, the conversation moves forward to the next dialogue step, and a new output context is created, which then becomes the input context for the following step in the conversation. If the user's input does not match any intent, the fallback intent is triggered instead. Even then, the same output context is generated, so the conversation still progresses to the next step in the intended order.

An example conversation is provided here. The chatbot begins the dialogue with a greeting and a question. At this point, the output context from the previous dialogue, labeled "1-followup," is active. The intents "2-yes," "2-no," and "2-fallback" all have "1-followup" as their input context, meaning that any of these three intents can be recognized based on the user's response.

If the user responds positively, the "2-yes" intent is activated; if the response is negative, the "2-no" intent is activated. If the user's response doesn't match either of these intents, the "2-fallback" intent is activated. The chatbot then replies with the specific response coded for each intent. All three of these intents—"2-yes," "2-no,"





**Figure 3.2:** Dialog flow of all the chatbot

and "2-fallback"—activate the "2-followup" context, which then serves as the input context for the next set of intents in the conversation.

An example where 2-yes intent is activated.

**Chatbot** Hello, I'm Travelbot. I have a deep passion for traveling and enjoy engaging in conversations about it. What are your thoughts on traveling?

**User** I like traveling, it is relaxing and fun!

**Chatbot** Awesome. I can't wait for us to share our amazing travel experiences. Where was your last journey to?

Example where 2-no intent is activated.

**Chatbot** Hello, I'm Travelbot. I have a deep passion for traveling and enjoy engaging in conversations about it. What are your thoughts on traveling?

**User** Not really.

**Chatbot** No worries at all. Let's share our travel experiences; they might inspire a newfound love for traveling. Where was your last journey to?

Example where 2-fallback intent is activated.

**Chatbot** Hello, I'm Travelbot. I have a deep passion for traveling and enjoy engaging in conversations about it. What are your thoughts on traveling?

**User** Sometimes yes, sometimes no. It's a toss up

**Chatbot** I can't wait for us to share amazing travel experiences. Where was your last journey to?

### 3.3 Integration of Linguistic Features

In this experiment, three distinct chatbots were developed to explore the effects of different linguistic features on user interaction: a chatbot incorporated with intensifiers, a chatbot incorporated with emotional language, and a chatbot which had neither of these features. All three chatbots were built on the same foundational dialogue structure, ensuring that any variations in user response could be attributed to specific linguistic modifications rather than differences in content.

#### 3.3.1 Control Chatbot

Firstly the control chatbot was designed. For this, as stated in the subsection above, open-ended questions and responses were designed manually. The control chatbot was designed without incorporating intensifiers or emotional language. The aim of designing its dialogue was to ensure it remained low in emotional expression as well as to reduce the use of intensifiers, providing a baseline for comparison in the study.

#### 3.3.2 Intensifier Chatbot

Intensive adverbs, also known as intensifiers, are words that modify the degree or intensity of the action or quality being described. Examples of such intensifiers include "really," "so," "very," "awfully," "terribly," "certainly," "absolutely," and "totally."

In this study, the chatbot that incorporated intensifiers had 50 instances of these adverbs throughout its dialogues compared to zero in the other two. To ensure consistency across the dialogues, the content of the control and the intensifier chatbot

was kept exactly the same, with the sole difference being the presence of intensifiers in the dialogues of the latter. The dialogues were created for the control chatbot, and ChatGPT was then used to modify them. Specifically, the dialogues from the control chatbot were provided to ChatGPT, and it was prompted to insert intensifiers into the dialogue and generate various versions incorporating these intensifiers. In addition to the automated process, intensifiers were also manually added to the dialogue to ensure that this feature was as prevalent as possible without making the dialogue sound unnatural.

Example of dialogues for the control chatbot and intensifier chatbot.

<b>Control</b>	Interesting. That appears to be a suitable vacation spot. Personally, I appreciate cities that blend modernity with tradition and have a rich history. Do you also prefer cities with those qualities?
<b>Chatbot</b>	Personally, I appreciate cities that blend modernity with tradition and have an exceptionally rich history. Do you also prefer cities with those qualities?
<b>Intensifier</b>	So interesting. That appears to be quite a suitable vacation spot. Personally, I appreciate cities that blend modernity with tradition and have an exceptionally rich history. Do you also prefer cities with those qualities?
<b>Chatbot</b>	Personally, I appreciate cities that blend modernity with tradition and have an exceptionally rich history. Do you also prefer cities with those qualities?

### 3.3.3 Emotional Chatbot

Emotional language and emotion-inducing language involve mentions of an emotion or feeling, such as “happy,” “enticing,” or “depressing.” They also include words or phrases that induce emotion in the reader. Given the vague nature of this concept, several methods were employed in the design of the chatbot to ensure representation of the feature like in the studies that found them.

#### Verifying Emotional Content Using LIWC and VADER

In several studies that have identified references to emotion, independent coders were used to code these features [22] [19]. Consequently, no specific information is available from those studies about how they had coded emotional language and what was the threshold used. However, in other research papers, tools like LIWC (Linguistic Inquiry and Word Count), a textual analysis program, were utilized [40] [24] [43]. LIWC analyzes text samples on a word-by-word basis, comparing each word to an ever-evolving dictionary. There are word categories that are predefined groups that classify words based on their linguistic and psychological characteristics. These categories allow for the systematic analysis of language by organizing words into meaningful groups such as emotions, cognitive processes, and social interactions. The study by Newman et al. [40] used the LIWC 2007 dictionary for their analysis, and the output was represented as a percentage of the total

words in the sample.

In this study, LIWC was used both for the manipulation check and during the creation of the chatbot to ensure that the emotional chatbot contained the necessary features. The LIWC software was used to compare dialogues during their creation to achieve a high score for word category counts. This score represents the percentage of words in the text that belong to the particular category. For the design of the emotional chatbot, the categories that were used for comparison are the ones related to emotions like overall emotions (emo), positive emotions (emo\_pos), negative emotions (emo\_neg), anxiety (emo\_anx), anger (emo\_ang), and sadness (emo\_sad). However, since the chatbot is designed to focus on topics like travel and casual conversation, emotions such as anxiety and anger are not included in any of the chatbot's responses. Therefore these categories are taken out. The LIWC analysis of all the final versions of the chatbot for the relevant categories is presented in table 3.5. The Test Kitchen corpus is a comprehensive collection of English language samples, designed to reflect a wide range of writing and speaking contexts [53]. It consists of 15 different data sets, totaling 31 million words. Scores for different LIWC categories were calculated in this corpus. These scores which can be thought of an average, are also provided in the table for comparison. The overall emotion average for the Test Kitchen corpus is 1.88, with the Control chatbot scoring slightly lower at 1.8 and the Intensifier chatbot at 2.11. The Emotional chatbot significantly surpasses both, scoring 5.47, indicating a higher emotional tone. For positive emotions, the Test Kitchen corpus average is 1.10, while the Control and Intensifier chatbots score 1.8 and 1.83, respectively. The Emotional chatbot shows a strong emphasis on positive expression with a score of 4.45. This is possibly due to the nature of chatbot, engaging in casual and positive conversation. In negative emotions, the corpus average is 0.65; the Control chatbot scores 0, while the Intensifier and Emotional chatbots score 0.28 and 0.3, respectively. Both the Control and Intensifier chatbots show no sadness which is lesser than the corpus score of 0.13, with the Emotional chatbot registering a score of 0.2, a slightly higher score. Overall, these findings highlight that the Emotional chatbot displays higher emotional content compared to the Control chatbot, particularly in overall and positive emotions.

LIWC does not account for sentiment-bearing lexical items such as acronyms and slang, which are crucial for analyzing text. Additionally, it cannot differentiate the intensity of sentiment in words. For example, "The food was okay" and "The food was exceptional" would score equally, as both sentences contain a positive term.

To address these limitations, we also used VADER (Valence Aware Dictionary for sEntiment Reasoning), a sentiment analysis tool designed specifically for social media contexts. VADER uses both qualitative and quantitative methods to create a sentiment lexicon, which was then empirically validated by human annotators and

	<b>Test kitchen corpus</b>	<b>Control</b>	<b>Intensifier</b>	<b>Emotional</b>
emo	1.88	1.8	2.11	5.47
emo <sub>pos</sub>	1.10	1.8	1.83	4.45
emo <sub>neg</sub>	0.65	0	0.28	0.3
emo <sub>sad</sub>	0.13	0	0	0.2

**Table 3.5:** LIWC Analysis Scores for All Dialogues Created for Each Chatbot

with performance measures to ensure high accuracy [54]. Unlike LIWC, VADER accounts for sentiment intensity and provides a normalized score from -1 to 1, where -1 is negative and 1 is positive. It also provides scores for what proportion of the text is a certain emotion (positive, negative, and neutral). The VADER lexicon contains words and their sentiment scores, which we manually used to add words with high emotions to the chatbot’s dialogue [54].

A VADER sentiment analysis was conducted to confirm that the emotion expressed by the chatbot was stronger than that of both the control chatbot and the intensifier chatbot. The VADER sentiment scores are presented in table 3.6.

### Designing the Emotional Chatbot Dialogues

Similar to the intensifier chatbot, dialogues for the control chatbot were initially created for the creation of the emotional chatbot. These dialogues were then provided to ChatGPT, which was prompted to “make the dialogues emotional or emotion-inducing,” resulting in various modified versions.

In addition to this, the emotional content was further enhanced by manually adding words from the VADER lexicon to the dialogues, ensuring that this feature was as prevalent as possible without making the dialogue sound unnatural. To verify that the emotional chatbot had a higher emotional score compared to the others, periodic checks were performed using the LIWC software.

<b>File</b>	<b>control chatbot</b>	<b>emotional chatbot</b>	<b>intensifier chatbot</b>
<b>Negative</b>	0.012	0.018	0.013
<b>Neutral</b>	0.771	0.687	0.766
<b>Positive</b>	0.217	0.295	0.212

**Table 3.6:** Sentiment Scores for All Dialogues Created for Each Chatbot

Example of dialogues for the control chatbot and emotional chatbot.

- Control** Interesting. That appears to be a suitable vacation spot. Personally, I appreciate cities that blend modernity with tradition and have a rich history. Do you also prefer cities with those qualities?
- Chatbot** So interesting. That appears to be quite a suitable vacation spot.
- Interesting** Personally, I appreciate cities that blend modernity with tradition and have an exceptionally rich history. Do you also prefer cities with those qualities?
- Chatbot** Oh, wonderful! That sounds like such a delightful destination. I'm particularly drawn to cities steeped in history, where every corner tells a story of traditions. Do you also enjoy exploring cities?
- Emotional**
- Chatbot**

Examples of full conversation with all the chatbots are provided in appendix B.

# Research Methodology

The purpose of this study was to assess if linguistic features affect the perception of gender in text-based conversational agents or chatbots. The study aimed to address the following research questions and hypotheses.

**How does the use of linguistic features, specifically intensifiers and emotion-inducing words, influence the perception of gender in text-based conversational agents?**

- *H1: The usage of intensifiers by the chatbot results in increased perception of the chatbot as feminine*
- *H2: The usage of emotion-inducing words by the chatbot results in increased perception of the chatbot as feminine*

**How does the effect of intensifiers on the perceived gender of text-based conversational agents compare to the effect of emotion-inducing words?**

To investigate these questions, chatbots incorporated with these features were created and participants interacted with them. This chapter will provide detailed information on the research design, participants, procedure, measures used, and data analysis methods.

## 4.1 Experimental Design

The experiment involved three different chatbots: one incorporating intensifiers, one using emotional language, and one with neither feature. A between-subjects design was employed, with participants interacting with only one of the chatbots.

Several reasons justified the choice of a between-subjects design. Firstly, it helped reduce familiarity bias. Interacting with three similar chatbots could lead

to carryover effects, where the experience of one condition affects the response to subsequent conditions due to familiarity with the experimental setup or stimuli. Also, this design minimized the risk of participant fatigue. Another key consideration was that repeated exposure to the survey could introduce biases in future interactions and responses.

## 4.2 Participant Recruitment

Participants were recruited through convenience sampling. The survey link was shared with various groups, directing participants to the study. Participants remained anonymous and did not need to provide any personal details, ensuring privacy and reducing social desirability bias [55]. This approach facilitated a diverse range of participants while maintaining their confidentiality.

The experiment was conducted over three weeks with a total of 132 participants across three different conditions. Specifically, 58 participants interacted with a chatbot that used emotional language, 39 engaged with a chatbot that employed intensifiers, and 39 interacted with a control chatbot. While all of them participated in the interaction, only 92 of them successfully completed the conversation. The chatbot dialogue has a code in its final dialogue. Participants were asked to provide this code in the survey. This was to confirm that the participants completed the conversation and had adequate exposure to the manipulations. The responses with the code are the ones included in the analysis. Ultimately, we obtained usable data from 31 participants in the control chatbot condition, 31 in the intensifier chatbot condition, and 30 in the emotional language chatbot condition. The number of participants who completed the interaction with the emotional language chatbot and provided the code was only 53%, which is why the number of participants in this condition was increased to ensure adequate data collection, to prevent uneven participant numbers across conditions.

### Age Distribution

Participants were divided into six age groups. The largest age group was 18-24 years, comprising 29 participants (31.5% of the total). The 25-34 years age group included 24 participants (26.1%). The 35-44 years age group had 16 participants (17.4%). The 45-54 years age group comprised 13 participants (14.1%). The 55-64 years age group had 9 participants (9.8%). Finally, the 65 or older age group had just 1 participant (1.1%). The table providing detailed demographic data on the number of participants in each age group who interacted with each chatbot is present in the appendixD.1. It is also illustrated in the4.1a.



### **Gender Distribution**

The gender distribution of the participants was relatively balanced. Out of the 92 participants, 44 identified as men (47.8%) and 48 identified as women (52.2%). Among the men, 14 interacted with the control chatbot, 16 with the intensifier chatbot, and 14 with the emotional chatbot. Among the women, 17 interacted with the control chatbot, 15 with the intensifier chatbot, and 16 with the emotional chatbot. This is illustrated in the table D.1 and figure below 4.1b.

## **4.3 Survey design and Measurements**

The data collected included both a detailed survey and a log of the interactions with the chatbot. The survey featured a mix of quantitative and qualitative questions, aimed at gaining a comprehensive understanding of participants' interactions and perceptions. The questions, along with their options and scales, are provided in the table below 4.1. However, as the survey was anonymous, the interaction logs and their corresponding survey responses cannot be matched for participants. Despite this limitation, the overall analysis can provide valuable insights into how participants experienced and interpreted their conversations with the chatbot.

The survey was designed with several key aims: to assess participants' perceptions of the chatbot's gender, verify the effectiveness of the linguistic manipulations, and evaluate various dimensions of the overall interaction experience. The survey design is detailed below.

### **4.3.1 Demographics**

Participants were asked to provide basic demographic information, including age and gender. This data was used to understand the diversity of the participant pool and to analyze any demographic-related trends. The study involved a total of 92 participants who interacted with three different chatbots: control, intensifier, and emotional chatbots. The detailed demographic information is illustrated in the tables provided in appendix D.

### **4.3.2 Experience with Chatbots**

To capture participants' familiarity with chatbots, a question about their previous interactions with chatbots was included. Participants were asked how frequently they used chatbots. This information helped in understanding the baseline familiarity and comfort level of participants with chatbot technology. Participants' prior experience

with chatbots varied. Fifteen participants (16.3%) had never talked to a chatbot before. Eight participants (8.7%) reported talking to chatbots about once a month. Eighteen participants (19.6%) indicated that they talk to chatbots frequently, almost every day. The largest group, comprising 51 participants (55.4%), had talked to a chatbot a few times but not often. The tables provide detailed demographic data on the number of participants in each group who interacted with each chatbot is present in the appendix D.1 which is also illustrated in the figure 4.1c.

### 4.3.3 Gender Perception

Several methods were employed to assess participants' perceptions of the chatbot's gender. Directly asking participants to identify the chatbot's gender might invoke bias. Instead, participants were asked to rank the masculinity, femininity, and gender neutrality of the chatbot based on their interaction on a scale of 1 to 7. According to Bem's Sex-Role Inventory (BSRI), femininity and masculinity should be rated as separate dimensions rather than as extremes of the same scale [56]. This approach allows for the characterization of agents as more feminine (high femininity / low masculinity), masculine (high masculinity / low femininity), androgynous (high femininity / high masculinity), or undifferentiated (low femininity / low masculinity). A gender neutrality scale was used to provide a nuanced perspective of the chatbot's gender perception [57].

Bem sex role inventory (BSRI), a 60-item survey, is a method used to characterize a person as masculine, feminine, or androgynous based on their endorsement of masculine and feminine personality characteristics. A short form of the BSRI was adapted for the context of a brief conversation with a chatbot about travel. Items not relevant to the conversation, such as "loves children," were excluded [56]. In the adapted set of questions, 5 masculine and 5 feminine characteristics are present which are used to calculate masculinity and femininity scores. The five feminine characteristics selected were Gentle, Sympathetic, Sensitive to needs of others, Warm, and Understanding. The five selected masculine traits were Aggressive, Dominant, Forceful, and Strong personality. Participants were asked to rate these on a scale of 1 to 7.

Participants were also asked to give the chatbot a human name, providing an indirect measure of gender assignment. This open-ended question allowed participants to project their perceptions without being directly prompted about gender. Additionally, participants were asked to describe the chatbot. This qualitative response could reveal gender perceptions through the use of pronouns and descriptive terms that might indicate perceived gender.

### 4.3.4 Assessing the manipulations

To ensure the effectiveness of the chatbot manipulations—linguistic features such as intensifiers and emotion-inducing language—several questions were developed to assess participants' awareness of these features. Participants were asked if they noticed any linguistic features in their conversation. This multiple-choice question included distractors alongside the actual linguistic manipulations, such as questions, personal pronouns, and uncertainty verbs.

Participants were also asked if they found the chatbot enthusiastic, expressive, joyful, or emotional. "Enthusiastic" and "expressive" could be related to high usage of intensifiers, while "joyful" and "emotional" could be related to the detection of emotional language in the chatbot. The description of the chatbot provided by participants also served as an indicator of their perception of the linguistic features used.

### 4.3.5 Anthropomorphism

Anthropomorphism is closely related to gender. Conversational agents and social robots are often designed with specific genders to increase human likeness. Borau et al. [12] found that female chatbots are perceived as more human than male counterparts when compared with non-human entities like animals and machines. Thus, anthropomorphism might affect gender perception. To analyze this, two questions from the Godspeed questionnaire were used, assessing the dimensions of fake vs. natural and humanlike vs. machinelike which the participants rated on a scale of 1 to 7. These questions helped gauge the extent to which participants perceived the chatbot as human-like, which could, in turn, influence their gender perception.

## 4.4 Procedure

The experiment primarily consisted of the participant interacting with one of the chatbots and subsequently answering a survey. The experiment was conducted online in the following steps. Participants were first introduced to the experiment through an information sheet, detailing what the study entailed. However, to prevent social desirability biases, the true purpose of the study was not disclosed until the end. After reviewing the experiment's details, participants provided their consent to participate. They were informed that they could withdraw from the study at any time until the survey was submitted. Once consent was given, participants engaged in an interaction with a chatbot. Following this interaction, they completed a series

of qualitative and quantitative questions designed to assess their experience and address the research questions.

Upon completing the survey, participants were debriefed on the actual purpose of the study. The information sheet, consent form, and debriefing document are all included in the appendix C.

## 4.5 Data analysis

Data analysis was conducted using SPSS, focusing on both descriptive and inferential statistics. Descriptive statistics were computed for all survey measures. For gender perception measures, including the perception of masculinity, femininity, and gender neutrality, as well as BSRI masculine and feminine scores, ANOVA was performed across all chatbot groups. To

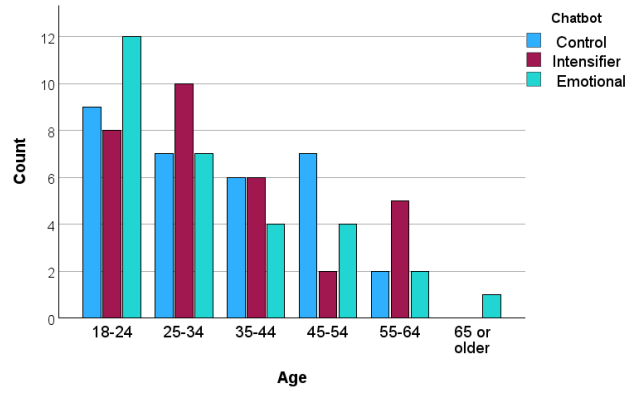
For other measures, descriptive statistics such as median, mean, and standard deviation were utilized. Due to the multicultural nature of the names (Indian and European), manual categorization was employed for names and descriptions. Descriptions were analyzed both manually and with Python NLP techniques to identify pronouns and adjectives.

*Table 4.1: Survey questions and the scale along with their objectives.*

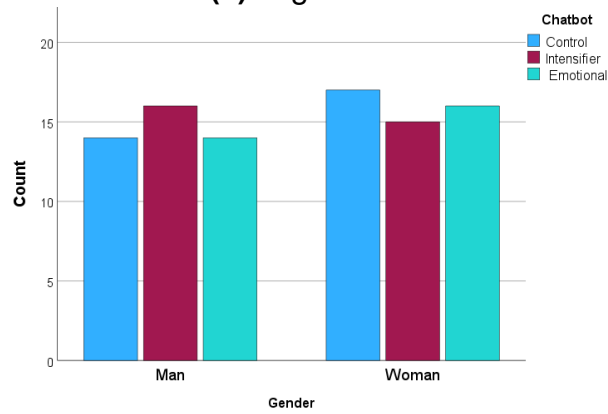
<b>Objective</b>	<b>Questions</b>	<b>Type</b>
Demographics	Please specify your age range: - 18-24 - 25-34 - 35-44 - 45-54 - 55-64 - 65 or older	Single Choice
	Please select your gender: - Woman - Man - Non-binary	Single Choice
Experience with Chatbots	How often do you interact with chatbots? - I have never talked to a chatbot before. - I've talked to a chatbot a few times, but not often. - I talk to chatbots about once a month or so. - I talk to chatbots frequently, almost every day.	Single Choice

<b>Objective</b>	<b>Questions</b>	<b>Type</b>
Gender perception	<p>On a scale from 1 to 7, where 1 means "Not at all" and 7 means "Very," how would you rate the chatbot in terms of the following characteristics?</p> <ul style="list-style-type: none"> <li>- Assertive</li> <li>- Understanding</li> <li>- Warm</li> <li>- Strong personality</li> <li>- Forceful</li> <li>- Dominant</li> <li>- Sympathetic</li> <li>- Sensitive to needs of others</li> <li>- Aggressive</li> <li>- Gentle</li> </ul>	Likert Scale
	<p>On a scale from 1 to 7, where 1 means "Not at all" and 7 means "Very," how would you rate the chatbot in terms of the following?</p> <ul style="list-style-type: none"> <li>- Masculine</li> <li>- Feminine</li> <li>- Gender-Neutral</li> </ul>	Likert Scale
Questions Covering Multiple Constructs	Suggest a human name that best fits the chatbot you just interacted with.	Open-Ended
	In a sentence or two, describe the chatbot from your interaction.	Open-Ended
Manipulation Check	<p>On a scale from 1 to 7, where 1 means "Not at all" and 7 means "Very much," how would you rate the chatbot in terms of the following characteristics?</p> <ul style="list-style-type: none"> <li>- Enthusiastic</li> <li>- Emotional</li> <li>- Expressive</li> <li>- Joyful</li> </ul>	Likert Scale

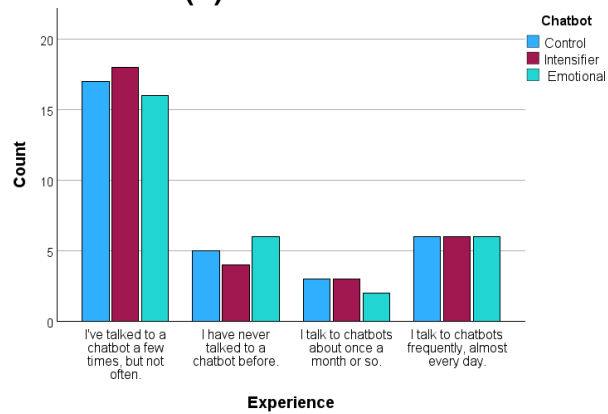
Objective	Questions	Type
	<p>Which of the following features stood out to you in the language used by the chatbot? (Select multiple answers)</p> <ul style="list-style-type: none"> <li>- Personal pronouns (e.g., I, you, we)</li> <li>- Uncertainty verbs (e.g., "might help you," "can be tricky")</li> <li>- Intensifiers (e.g., so, very, totally, absolutely)</li> <li>- Questions (e.g., "What are your thoughts on...?", "Do you like...?")</li> <li>- Emotional language (e.g., "Wow, that's fantastic", "deep passion", "miss you")</li> <li>- Fillers (e.g., "Oh", "Ahh", "like")</li> </ul>	Multiple Choice
Anthropomorphism	On a scale from 1 to 7, where 1 means "Fake" and 7 means "Natural," how would you rate the chatbot?	Likert Scale
	On a scale from 1 to 7, where 1 means "Machine-like" and 7 means "Humanlike," how would you describe the chatbot?	Likert Scale



(a): Age distribution



(b): Gender distribution



(c): Experience distribution

Figure 4.1: Demographic and Experience distribution across chatbots.

# Results

This section presents the findings of the experiment, detailing the measures employed to address the research questions and explore the interaction in depth. The results are illustrated and explained in the following subsections.

## 5.1 Perceptions of Chatbot Gender

To assess participants' perceptions of the chatbot's gender, as outlined in the previous chapter, the survey included several questions. We evaluated perceived masculinity, femininity, and gender neutrality, along with ratings from the Short Form of the Bem Sex Role Inventory. Additionally, we analyzed the names and descriptions participants provided for the chatbots. The detailed results are presented below.

### 5.1.1 Perceived masculinity, femininity and gender neutrality

Variable	Chatbot	N	Mean	Std. Deviation	Std. Error Mean
Masculine	Control	31	3.23	1.98	0.35
	Intensifier	31	2.84	1.81	0.32
	Emotional	30	2.90	1.86	0.34
Feminine	Control	31	3.29	1.97	0.35
	Intensifier	31	4.16	2.08	0.37
	Emotional	30	3.70	2.05	0.37
Gender Neutrality	Control	31	4.35	2.42	0.43
	Intensifier	31	4.81	1.80	0.32
	Emotional	30	4.33	2.02	0.37

**Table 5.1:** Descriptive statistics of Perceived Masculinity, Femininity and Gender Neutrality



Variable	Chatbot groups	Mean difference
Masculine	Control and Intensifier	0.39
	Control and Emotional	0.33
	Intensifier and Emotional	0.06
Feminine	Control and Intensifier	0.87
	Control and Emotional	0.41
	Intensifier and Emotional	0.46
Gender Neutrality	Control and Intensifier	0.45
	Control and Emotional	0.02
	Intensifier and Emotional	0.47

**Table 5.2:** Mean differences between the groups

Measure	F	p-value	Effect Size (Eta-squared)
<b>Masculine</b>	0.377	0.687	0.008
<b>Feminine</b>	1.420	0.247	0.031
<b>Gender Neutrality</b>	0.501	0.608	0.011

**Table 5.3:** ANOVA Summary for Perceived Masculinity, Femininity, and Gender Neutrality

### Masculinity

Participants rated the perceived masculinity of the chatbots on a scale from 1 (Not Masculine) to 7 (Very Masculine). The control chatbot was perceived as the most masculine (Mean = 3.23, SD = 1.98), while the emotional chatbot was rated marginally lower (Mean = 2.90, SD = 1.86) and the intensifier chatbot had the lowest score (Mean = 2.84, SD = 1.81). The standard deviations indicate moderate variability in ratings across all chatbot conditions.

To determine if there are significant differences between the perceived masculinity across three chatbots, a one-way ANOVA was done. The ANOVA results for perceived masculinity were not significant ( $F(2, 89) = 0.377, p = 0.687$ ). The mean differences between the chatbot pairs indicate that the difference in the measure between the control and intensifier chatbots is greater than that between the control and emotional chatbots. This observation might suggest that the intensifier has a higher impact compared to emotion-inducing words. However, due to the lack of statistical significance, this should be interpreted as a potential trend rather than a confirmed effect. Table 5.1 provides the descriptive statistics, table 5.2 presents the mean difference of the measure between groups, and table 5.3 presents the ANOVA results of the measure.

## Femininity

For perceived femininity rated on the same scale, the intensifier chatbot was rated as the most feminine (Mean = 4.16, SD = 2.08), while the control chatbot was rated the least (Mean = 3.29, SD = 1.97) and the emotional chatbot's femininity rating fell between the other two, with a mean of 3.70 (SD = 2.05). These ratings suggest that the intensifier chatbot was perceived as more feminine compared to the control chatbot and the emotional chatbot.

A one-way ANOVA was done to determine if there are significant differences between the perceived femininity across three chatbots. The ANOVA for perceived femininity also did not yield significant results ( $F(2, 89) = 1.420, p = 0.247$ ). Similar to perceived masculinity, the mean differences between the chatbot pairs indicate that the difference in the measure between the control and intensifier chatbots is greater than that between the control and emotional chatbots. However, the difference is higher in this case. This observation might suggest that the intensifier has a higher impact compared to emotion-inducing words. However, due to the lack of statistical significance, this should also be interpreted as a potential trend rather than a confirmed effect.

## Gender Neutrality

Regarding gender neutrality, the intensifier chatbot was rated as the most gender-neutral (Mean = 4.81, SD = 1.80), while the control chatbot's mean score was lower (Mean = 4.35, SD = 2.42) and the emotional chatbot had the lowest mean (Mean = 4.33, SD = 2.02) among the three. The standard deviations indicate some variability in the perceived gender neutrality of the chatbots. For the Gender Neutrality measure, the highest mean difference is between the intensifier and emotional chatbots. The ANOVA for gender neutrality was not significant ( $F(2, 89) = 0.501, p = 0.608$ ) indicating that there are no significant differences between the groups.

### 5.1.2 Short form Bem Sex Role Inventory

To assess the perceived femininity and masculinity of each chatbot, scores were derived from the Bem Sex Role Inventory (BSRI). Each score was calculated by averaging responses from 10 features: 5 related to femininity and 5 related to masculinity. The average of these features provided a mean femininity and masculinity score for each chatbot. Table 5.4 provides the descriptive statistics, table 5.5 presents the mean difference of the measure between groups, and table ?? presents the ANOVA results of the measures.

Variable	Chatbot	N	Mean	Std. Deviation	Std. Error Mean
Femininity BSRI	Control	31	4.36	1.55	0.28
	Intensifier	31	4.90	1.70	0.31
	Emotional	30	4.70	1.28	0.23
Masculinity BSRI	Control	31	3.61	1.61	0.29
	Intensifier	31	3.46	1.45	0.26
	Emotional	30	3.57	1.29	0.24

**Table 5.4:** Descriptive statistics for BSRI scores for masculinity and femininity among all chatbots

Variable	Chatbot groups	Mean difference
Femininity-BSRI	Control and Intensifier	0.53
	Control and Emotional	0.34
	Intensifier and Emotional	0.20
Masculinity-BSRI	Control and Intensifier	0.15
	Control and Emotional	0.04
	Intensifier and Emotional	0.11

**Table 5.5:** Mean differences between the groups

### Masculinity score

The internal consistency of the masculinity sub-scale was  $\alpha = 0.817$ , indicating good reliability. Although removing "Strong Personality", one among the 5 items, would slightly increase Cronbach's Alpha to 0.841, this improvement is minimal. Therefore, all five items are retained, and the mean of these items is calculated to determine the masculinity score.

On the masculinity scale, the Control Chatbot had a slightly higher mean score of 3.61 (SD = 1.62) compared to the mean score of 3.57 (SD = 1.29) for the Emotional Chatbot and 3.46 (SD = 1.45) for the Intensifier Chatbot. This suggests that the Control Chatbot was perceived as slightly more masculine than the others. A separate ANOVA was performed to assess differences in masculinity scores across the groups. The analysis showed that the differences were not statistically significant, ( $F(2, 89) = 0.09, p = 0.92$ ). The mean difference between the control and intensifier chatbots is greater than the difference between the control and emotional chatbots. However, this difference is minimal. Due to the lack of statistical significance, this should be interpreted as a minor trend rather than a confirmed effect.

## Femininity Score

The internal consistency of the femininity sub-scale was  $\alpha = 0.907$ , reflecting high reliability. All five items are retained for calculating the mean femininity score.

For femininity, the Control Chatbot had a mean score of 4.36 (SD = 1.55), which was lower than the mean score of the Emotional Chatbot 4.70 (SD = 1.28), and the Intensifier Chatbot's mean score of 4.90 (SD = 1.704). This suggests that the Intensifier Chatbot was perceived as most feminine out of all three.

An ANOVA was conducted to examine the differences in femininity scores across the groups. The results indicated that there were no statistically significant differences in femininity scores between the groups (  $F(2, 89) = 0.98, p = 0.38$  ).

The mean difference between the control and intensifier chatbots is greater than the difference between the control and emotional chatbots. This suggests that the intensifier might have more impact on femininity compared to emotion-inducing words. Due to the absence of statistical significance, these findings are better viewed as indicative of a possible pattern or trend rather than a conclusive effect.

**Table 5.6:** ANOVA Results and Effect Sizes for Femininity and Masculinity (BSRI)

Measure	F	Sig.	Eta-squared
Femininity (BSRI)	0.98	0.38	0.02
Masculinity (BSRI)	0.09	0.92	0.00

## 5.2 Anthropomorphism

The analysis of anthropomorphism across three different chatbot conditions—control, intensifier, and emotional—revealed differences in how the chatbots were evaluated on the variables of perceived naturalness and humanlikeness.

### 5.2.1 Perceived Naturalness

Participants rated the naturalness of the chatbots on a scale from 1 (Fake) to 7 (Natural). The mean ratings for the control, intensifier, and emotional chatbots were 4.65, 4.35, and 4.33, respectively. The control chatbot was perceived as slightly more natural (Mean = 4.65, SD = 2.07) compared to the intensifier chatbot (Mean = 4.35, SD = 2.07) and the emotional chatbot (Mean = 4.33, SD = 1.65). The emotional chatbot, showed a slightly lower standard deviation (SD = 1.65), indicating less variability in user ratings.

An ANOVA was conducted to examine statistically significant differences and it indicated no significant difference between groups,  $F(2, 89) = 0.247, p = 0.782$ .

### 5.2.2 Perceived Humanlikeness

When examining the humanlikeness of the chatbots, rated on a scale from 1 (Machinelike) to 7 (Humanlike), the intensifier chatbot (Mean = 4.32, SD = 2.04) received a moderately higher mean score compared to the control chatbot (Mean = 4.26, SD = 2.16). Notably, the emotional chatbot was rated the lowest in terms of humanlikeness (Mean = 3.53, SD = 1.78).

Similarly, an ANOVA was conducted to examine statistically significant differences and this dimension also showed no significant difference between groups, ( $F(2, 89) = 1.455$ ), ( $p = 0.239$ ).

These results suggest that the manipulations did not lead to significant differences in the perceptions of the chatbot on these dimensions. In conclusion, there is a trend of the control chatbot being rated higher in naturalness, while the intensifier chatbot was slightly more human-like. The emotional chatbot, however, received the lowest scores across both dimensions, suggesting that the added emotional elements may have impacted users' perceptions negatively in terms of anthropomorphism.

Variable	Chatbot	N	Mean	Std. Deviation	Std. Error Mean
Fake-1 Natural-7	Control	31	4.65	2.07	0.37
	Intensifier	31	4.35	2.07	0.37
	Emotional	30	4.33	1.65	0.30
Machinelike-1 Humanlike-7	Control	31	4.26	2.16	0.39
	Intensifier	31	4.32	2.04	0.37
	Emotional	30	3.53	1.78	0.32

**Table 5.7:** Descriptive Statistics for Measures Informing Anthropomorphism Across All Chatbots

Dependent Variable	F	Sig.	Eta-squared
Fake_Natural	0.247	0.782	0.006
Machinelike_Humanlike	1.455	0.239	0.032

**Table 5.8:** ANOVA Results and Effect Sizes for Perceived Naturalness and Humanlikeness

### 5.3 Relationship between Gender of the participants and Gender perception across all chatbots

Women reported higher femininity scores (Mean = 4.97, SD = 1.39) compared to men (Mean = 4.31, SD = 1.61), with a p-value of 0.038 in the ANOVA. A similar pattern emerged in the BSRI femininity scores, where women scored a higher bsri score (Mean = 5.15, SD = 1.44) to the chatbots than men (Mean = 4.41, SD = 1.64), with a p-value of 0.024. However, no significant differences were observed in masculinity scores, BSRI masculinity, or gender neutrality between men and women.

Measure	Group	Mean	Std. Deviation
<b>Masculinity</b>	Men	3.61	1.56
	Women	3.49	1.35
<b>Femininity</b>	Men	4.31	1.61
	Women	4.97	1.39
<b>Gender Neutrality</b>	Men	4.55	2.10
	Women	4.46	2.09
<b>BSRI Femininity</b>	Men	4.41	1.64
	Women	5.15	1.44
<b>BSRI Masculinity</b>	Men	3.48	1.63
	Women	3.11	1.53

**Table 5.9:** Descriptive Statistics for Masculine, Feminine, Gender Neutrality, BSRI Femininity, and BSRI Masculinity.

Measure	F	Sig.
<b>Masculine</b>	0.162	0.689
<b>Feminine</b>	4.425	0.038
<b>Gender Neutrality</b>	0.040	0.842
<b>BSRI Femininity</b>	5.244	0.024
<b>BSRI Masculinity</b>	1.213	0.274

**Table 5.10:** ANOVA Results for Masculine, Feminine, Gender Neutrality, BSRI Femininity, and BSRI Masculinity between genders.

For each chatbot, this relationship was explored, and it was found that for the intensifier chatbot, there was a significant difference in how women and men perceived it. The mean femininity score of men was 4.20 compared to a significantly higher 5.6 for women. An ANOVA of the measures yields significant differences between the groups.

Measure	Gender	N	Mean	Std. Deviation
femininitybsri	Men	16	4.20	1.93
	Women	15	5.64	1.03
Femininity	Men	16	4.69	1.70
	Women	15	3.60	2.35

**Table 5.11:** Group Statistics of the relationship between gender and gender perception in Intensifier chatbot between genders

Measure	F	Significance
femininitybsri	11.685	0.002
Feminine	4.697	0.039

**Table 5.12:** ANOVA Results

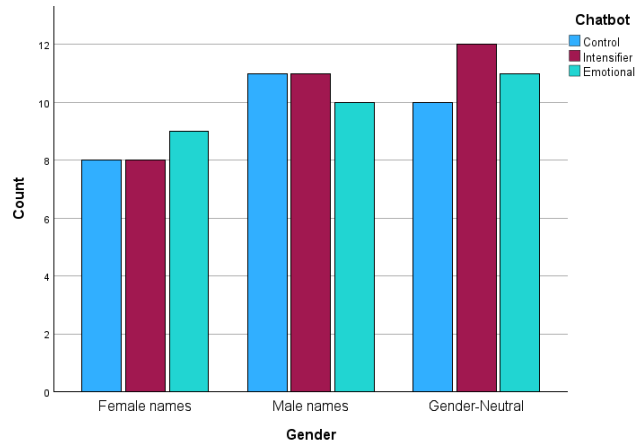
## 5.4 Results that deal with multiple measures-Gender perception and Anthropomorphism

Certain questions give insight into different measurements like gender perception and anthropomorphism at the same time.

### 5.4.1 Name assignment

The question of assigning a human name to the chatbots resulted in diverse responses. The figure shows the number of gendered (female and male) and non-gendered names given to each chatbot by participants. In total, participants assigned 31 female, 33 male, and 28 non-gendered names across all chatbots. Specifically, the control chatbot was named 8 female, 11 male, and 10 non-gendered or gender-neutral names. The intensifier chatbot received 8 female, 11 male, and 12 non-gendered names, while the emotional chatbot was given 9 female, 10 male, and 11 non-gendered names. These results contrast with the findings from the perceived scores and BSRI scores. Notably, the number of gendered and non-gendered names was almost the same across all chatbots. The intensifier chatbot had a higher number of male and gender-neutral names compared to female ones, which is contradictory to the perceived femininity and masculinity scores and BSRI scores. The data is illustrated in the figure 5.1.

This question also gave insight into the participant's anthropomorphism perception of each chatbot. As the highest number of non-gendered names were given to the intensifier chatbot, it could be presumed that this chatbot was perceived as less anthropomorphic. All the names given by the participants for the chatbots are



**Figure 5.1:** Number of female, male, and non-gendered names for each chatbot.

provided in appendix E.

## 5.4.2 Participant's Description of the Chatbot

### Usage of pronouns and gender perception

Participants were asked to describe the chatbot, with the expectation that this would lead to the use of pronouns. However, most descriptions did not include pronouns, as "chatbot" was often used instead. When pronouns were used, gender-neutral ones like "It" and "Its" were the most common. Specifically, 5 participants used gender-neutral pronouns to describe the control chatbot, 11 used it to describe the intensifier chatbot, and 6 used it to describe the emotional chatbot. Thus the intensifier chatbot was most frequently described using gender-neutral pronouns. Male pronouns were used only once each for the control and intensifier chatbots, and not at all for the emotional chatbot. Female pronouns were used only once, specifically for the intensifier chatbot.

### Thematic analysis of participants' descriptions of chatbots

Thematic analysis of participant comments about the chatbot revealed several key themes reflecting how users perceived its personality, responsiveness, and interaction quality.

- **The chatbot accurately understood and personalized responses:** Participants frequently noted that the chatbot could understand their input, with some stating that it "accurately picked up responses" and "understood the answers" they provided. The chatbot was also perceived as being able to engage with



user input in a personalized manner, with users feeling that their “input was appreciated and engaged with.”. These comments were seen more for emotional and intensifier chatbots compared to the control chatbot.

- **The chatbot did not understand or respond to user input:** Several participants expressed frustration over the chatbot’s inability to respond effectively to their input. Comments such as “didn’t understand my response,” “did not notice my answers,” and “not responding based on user answers” were common. Users often felt that the chatbot’s responses were pre-determined, describing them as “felt like preinstalled answers” or “answers seemed readymade,” which led to the perception that their input was ignored. Furthermore, some participants indicated that the chatbot’s interaction felt one-sided, mentioning that it “didn’t matter what I said” and “it is about the chatbot, not about the participant.” This gave the impression that the chatbot was more of a machine than a conversational partner, with some users stating it “didn’t listen to what I said” or “don’t care about my input.” These comments were seen across the board for all the chatbots.
- **The chatbot felt like a person with opinions and personality:** Several participants remarked that the chatbot felt human-like, attributing human traits to its behavior. Phrases like “felt like talking to a real person” and “seemed to have its own views like a human” were common. Some participants even mentioned that the chatbot expressed its own opinions, making it seem more than a simple machine. These comments were observed for the control and intensifier chatbot and were rarely observed for the emotional chatbot.
- **The chatbot was friendly and energetic:** Many participants described the chatbot as friendly, enthusiastic, and polite. Descriptors such as “lively,” “very bubbly,” and “understanding” were repeatedly used to characterize the chatbot’s personality. This energy and positive approach made the interaction feel pleasant for most users, with some mentioning that the chatbot “felt like a good friend.” The chatbot’s polite and energetic tone was seen as a significant contributor to the overall positive experience. These comments were observed in the descriptions of all chatbots.
- **The chatbot’s responses felt repetitive and limited:** Many participants felt that the chatbot’s responses were too repetitive and limited. Some described the chatbot as having “unnecessary repetition” and offering “limited responses,” which diminished the quality of the interaction. The fast response time often led users to feel that the chatbot’s replies were “too fast and thus artificial,” reducing the overall authenticity of the experience. These comments were seen

mainly for the control and intensifier chatbots.

- **The chatbot led the conversation in a forced or predictable way:** Participants expressed that the chatbot often led the conversation in a way that felt forced or overly predictable. For example, some felt that the chatbot “was expecting my answers to be a certain way” or “directed the conversation in a specific direction.” These comments reflect a lack of spontaneity, with users feeling that the chatbot sometimes forced the interaction, making it less dynamic and more pre-programmed. These comments were seen more for the emotional chatbot.
- **The chatbot lacked natural interaction and felt artificial:** While some participants appreciated the chatbot’s speed and energy, others described it as artificial or lacking in naturalness. Comments like “felt scripted” and “need to be more natural” were common, suggesting that, while the chatbot responded quickly, it didn’t always do so in a way that felt genuine. Many noted that the chatbot’s quick responses seemed automated, which made the interaction feel less authentic.
- **The chatbot was knowledgeable but lacked depth:** Participants acknowledged that the chatbot had a good knowledge base and could provide informative responses in certain areas. However, there was a consistent perception that the conversation lacked depth. Users felt that while the chatbot was “effective in its purpose” and “informative,” it could benefit from being “more intelligent” or offering more nuanced responses. The chatbot was sometimes described as “generic” and in need of “more dimensions” to engage users in more complex topics. This was also seen across the board.
- **The chatbot’s formal or impersonal tone made the conversation less engaging:** Several users found the chatbot’s tone too formal or impersonal for their liking. Descriptions such as “too formal” and “little out of touch” were common. This formality sometimes made the interaction feel less engaging, with participants stating that “it didn’t matter what I said” and that they felt the chatbot was not responding to their input in a meaningful way. This was seen more for the intensifier and emotional chatbots.

Intensifier Chatbot: Higher usage of gender-neutral pronouns to describe the intensifier chatbot suggests it might have been perceived as less anthropomorphic compared to the others. It was described using terms such as “artificial,” “one-way,” “machine,” and “programmed” more frequently than the other chatbots. The use of non-gendered pronouns and these specific adjectives indicates that this chatbot was

often seen as more mechanical or less human-like. The lower usage of gendered pronouns across all chatbots leaves this interpretation somewhat inconclusive.

**Emotional Chatbot:** This chatbot was often labeled as "friendly" and stood out for being more "talkative" compared to the other chatbots. It was also described as "energetic" once and as "good" on six occasions. However, the emotional chatbot was referred to as "readymade" and "programmed," suggesting it was perceived as less humanlike. These descriptors were used less frequently than similar terms applied to the intensifier chatbot. All the descriptions given by the participants for the chatbots are provided in the appendix E.

## **5.5 Assessment of Validity of Manipulation**

The chatbots were manipulated by incorporating specific features relevant to the research objectives. The Intensifier chatbot was designed to include a higher number of intensifiers compared to the Control and Emotional chatbots, while the Emotional chatbot was enriched with words and phrases that heightened the emotional intensity of the dialogues. To assess whether these manipulations were effective, participants who interacted with the chatbots were asked specific questions.

### **5.5.1 Feature Detection in Chatbots**

To determine if participants detected the manipulated features, they were asked which features stood out to them in the language used by the chatbot. Six options were provided, two of which were the actual features used for manipulation. The results are illustrated in Figure 5.2.

The use of intensifiers was noticed in the Emotional chatbot at a higher rate compared to the other two. Specifically, 15 participants identified the use of intensifiers in the Emotional chatbot, compared to only 7 participants who noticed these in the Control chatbot and 11 in the Intensifier chatbot. Additionally, 21 participants observed emotional language in the Intensifier chatbot, while only 14 participants detected this in the Control chatbot.

The Emotional chatbot, which was designed to include more emotionally charged language and phrases, was identified by participants as having a higher frequency of emotional language. A total of 23 participants noticed this feature in the Emotional chatbot, compared to 14 participants in the Control chatbot. However, emotional language was also noticed in the Intensifier chatbot despite it not including any such words or phrases.

The Chi-Square test was used to determine if differences in participants' recognition of features across chatbots are statistically significant or due to chance. The

analyses revealed the following results: for intensifiers, the p-value was 0.083, indicating a potential trend towards a relationship, though it did not reach conventional significance. For emotional language, the Chi-Square test yielded a p-value of 0.135, indicating no significant association either. The results of the test are presented in table 5.14

Overall, while there were some trends, none of the results reached conventional significance thresholds, indicating that chatbot manipulations do not significantly influence participants' recognition of features.

Feature	Control chatbot	Intensifier chatbot	Emotional chatbot
Questions	21	22	20
Intensifiers	7	11	15
Personal Pronouns	14	20	11
Fillers	4	11	8
Emotional Language	14	19	21
Uncertainty Verbs	4	5	8

**Table 5.13:** Number of people recognizing linguistic features across chatbots

Feature	Degrees of Freedom (df)	p-value
Uncertainty Verbs	2	0.35
Fillers	2	0.12
Questions	2	0.93
Intensifiers	2	0.08
Personal Pronouns	2	0.08
Emotional Language	2	0.13

**Table 5.14:** Chi-Square Test Results for Chatbot Features

### 5.5.2 Participant Perceptions of Chatbot Attributes

Participants rated the chatbots on four emotional attributes: Enthusiastic, Emotional, Expressive, and Joyful. These ratings were analyzed across three groups: Control (Group 0), Emotional (Group 1), and Intensifier (Group 2) chatbots. The results are summarized in Table 5.15.

**Enthusiastic:** The mean scores for enthusiasm were highest in the emotional chatbot ( $M = 5.90$ ,  $SD = 1.35$ ), followed closely by the intensifier chatbot ( $M = 5.74$ ,  $SD = 1.48$ ), with the control chatbot receiving the lowest ratings ( $M = 5.35$ ,  $SD = 1.62$ ). This pattern suggests a possible trend where participants perceived the emotional chatbot as slightly more enthusiastic. However, a one-way ANOVA showed

no statistically significant difference in enthusiasm scores between the groups (  $F(2, 89) = 1.085, p = 0.342$  ). The enthusiasm scores were relatively consistent across the conditions.

**Emotional:** Participants rated the intensifier chatbot higher on emotional intensity (  $M = 4.52, SD = 1.91$  ) compared to the emotional (  $M = 3.90, SD = 1.692$  ) and control chatbots (  $M = 3.94, SD = 1.73$  ). This may indicate a trend where the intensifier chatbot was perceived as more emotional. However, the one-way ANOVA results showed no statistically significant difference in emotional scores across the groups, (  $F(2, 89) = 1.162, p = 0.318$  ). The emotional responses did not show a significant variation between the conditions.

**Expressive:** The emotional chatbot received the highest mean score for expressiveness (  $M = 5.33, SD = 1.67$  ), followed by the intensifier chatbot (  $M = 5.10, SD = 1.78$  ), and the control chatbot (  $M = 4.97, SD = 1.68$  ). Although the emotional chatbot showed a slightly higher rating, a one-way ANOVA found no significant difference in expressiveness between the groups, (  $F(2, 89) = 0.357, p = 0.701$  ). This suggests that the expressiveness ratings were similar across all groups.

**Joyful:** Ratings for joyfulness were highest in the emotional chatbot (  $M = 5.80, SD = 1.16$  ), followed by the intensifier chatbot (  $M = 5.65, SD = 1.40$  ), with the control chatbot receiving the lowest ratings (  $M = 5.03, SD = 1.89$  ). This trend may suggest that participants viewed the emotional chatbot as more joyful. However, a one-way ANOVA showed no statistically significant difference in joyful scores between the groups, (  $F(2, 89) = 2.199, p = 0.117$  ).

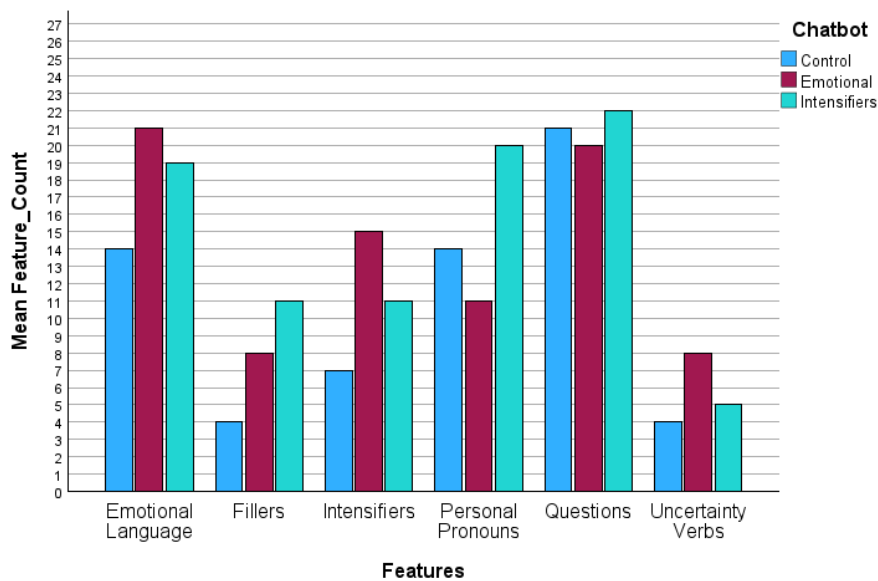
Overall, these results suggest some trends in how participants rated the chatbots, with the intensifier chatbot being perceived as more emotional and the emotional chatbot rated higher on other factors. However, none of these trends reached statistical significance. The Control chatbot consistently received lower ratings across all attributes, which aligns with its design as a neutral baseline.

Variable	Chatbot	N	Mean	Std. Deviation
<b>Enthusiastic</b>	Control	31	5.35	1.62
	Intensifier	31	5.74	1.48
	Emotional	30	5.90	1.35
<b>Emotional</b>	Control	31	3.94	1.69
	Intensifier	31	4.52	1.91
	Emotional	30	3.90	1.73
<b>Expressive</b>	Control	31	4.97	1.68
	Intensifier	31	5.10	1.78
	Emotional	30	5.33	1.67
<b>Joyful</b>	Control	31	5.03	1.89
	Intensifier	31	5.65	1.40
	Emotional	30	5.80	1.16

**Table 5.15:** Descriptive Statistics for Enthusiastic, Emotional, Expressive, and Joyful Characteristics Across Groups

Measurement	F	Sig.	Eta-squared
<b>Enthusiastic</b>	1.085	0.342	0.024
<b>Emotional</b>	1.162	0.318	0.025
<b>Expressive</b>	0.357	0.701	0.008
<b>Joyful</b>	2.199	0.117	0.047

**Table 5.16:** ANOVA Results with Effect Sizes



**Figure 5.2:** Features recognized by users for each chatbot

# Discussion and Conclusion

## 6.1 Discussion

### 6.1.1 Gender Perception

This study aimed to investigate whether linguistic features could influence the perception of gender in text-based conversational agents. Building on the research by Vanderlyn et al. [16], which demonstrated that linguistic styles—such as those characterized by empathy or practicality—could impact users' perceptions of a chatbot's gender and lead participants to project their own gender onto the chatbot, we sought to explore this phenomenon. Consequently, we identified key linguistic features to test whether these can affect the gender perception of a textual conversational agent.

After reviewing the literature on language use across different genders, we identified intensifiers and emotional language as prominent features present in female language. However, it remained unclear whether these features would similarly influence perceptions of gender in chatbots.

Therefore this research was guided by two primary questions:

**How does the use of linguistic features, specifically intensifiers and emotion-inducing words, influence the perception of gender in text-based conversational agents?**

**How does the effect of intensifiers on the perceived gender of text-based conversational agents compare to the effect of emotion-inducing words?**

Based on prior studies, the following hypotheses were made

**H1: The usage of intensifiers in the chatbot would result in increased per-**

## ception of the chatbot as feminine

### **H2: The usage of emotion-inducing words in the chatbot would result in increased perception of the chatbot as feminine**

The experiment revealed minor trends in the expected direction for some measures. Notably, participants perceived the intensifier chatbot as less masculine and more feminine compared to the control. This was reflected in both the perceived masculinity and femininity ratings, as well as the BSRI scores, with higher femininity and lower masculinity ratings for the intensifier chatbot. However, these results were not statistically significant, meaning they do not support the hypothesis that intensifiers can effectively influence the perception of a chatbot's gender. Nonetheless, it may be valuable to further explore how intensifiers, which are traditionally observed in female speech patterns, affect chatbot gender perceptions.

Similarly, the chatbot using emotion-inducing words showed lower perceived masculinity and higher perceived femininity, but the results were not statistically significant. While the descriptive statistics aligned with the hypothesis, the lack of significant findings suggests that emotion-inducing words do not have a proven effect on gender perception. Looking at the mean differences between the groups, the control-intensifier had a higher difference than the control-emotional for perceived femininity and masculinity as well as BSRI scores for femininity and masculinity. This says that intensifiers might have a higher impact compared to emotional words and emotion-inducing words.

The relationship between gender and gender perception is also interesting and worth looking into. Comparing the scores with the gender of the participants revealed that women scored the chatbots higher in femininity compared to men. This aligns with the result of the study where the participants projected their own gender to the chatbots. This is especially strong in the intensifier chatbot which showed significant results for the same. Women perceived the intensifier chatbot to have higher femininity and scored a higher BSRI femininity score compared to men. This means that they projected their own gender to the chatbot when intensifiers were used. This result was not observed for the other chatbots. Other factors, such as culture, nationality, and economic status, can influence perceptions of gender roles [58] [59]. These elements can also affect language use and its impact on gender perceptions, indicating a need for further exploration in this area. This aspect was not addressed in the present study, which included a mix of Indian and European participants. This diversity may have influenced the results regarding how gender and language were perceived.



### 6.1.2 Anthropomorphism

Another notable observation of the study is the relation between gender neutrality and anthropomorphism among the chatbots. The emotional chatbot was perceived as the most artificial and machine-like compared to the other two, and it also received the lowest rating for gender neutrality. This is in contrast to a previous study which found that gender-non-specific voices were rated lower in anthropomorphism [60]. This result was observed in a voice and not textual and not quite comparable. However, it is interesting to see that chatbots that were designed to have gendered features were seen as more artificial and machine-like.

## 6.2 Chatbot design implications

### 6.2.1 Feature integration

The analysis revealed that participants were most likely to notice the use of emotional language in the Emotional chatbot indicating that its emotional tone was prominent. Participants noticed this feature- usage of emotional language- at a higher rate in the Intensifiers chatbot too. The Control chatbot had the lowest recognition of emotional language, however, all the chatbots had a significantly high recognition of the feature. This could mean that the participants perceived all the chatbots as somewhat emotional. Creating a chatbot that engages in casual conversation about an informal topic like travel can lead to the character being presumed as slightly positive or cheerful. This was also seen through the thematic analysis where all the chatbots were described as fun and enthusiastic thus indicating they all were seen as somewhat emotional albeit positive

Participants were most likely to notice the use of intensifiers in the Emotional chatbot, which is surprising given that the Emotional chatbot was designed without any intensifiers. The Intensifier chatbot, on the other hand, also led to significant recognition of intensifiers, but to a lesser degree compared to the Emotional chatbot. This suggests that while intensifiers were present in the Intensifier chatbot, they were not as prominent. This could indicate that the use of intensifiers in the Intensifier chatbot was either too subtle or that additional intensifiers might have been needed to achieve higher visibility. However, feature detection might not be the most effective measure for a manipulation check, as language can be quite subtle, and specific features may not be easily noticeable.

Another way to detect the validity of manipulations was for participants to rate the chatbots on characteristics such as emotional, enthusiastic, joyful, and expressive, the intensifier chatbot was rated higher for being emotional. Conversely, the

emotional chatbot received lower ratings for emotional characteristics, even falling below the control chatbot. This result is unexpected but it might tell us that intensifiers can increase the emotional aspect of a text too. One possible explanation is that participants might have associated the term "emotional" with sadness or negative emotions. In contrast, the emotional chatbot was rated highest for joyfulness, expressiveness, and enthusiasm.

We had anticipated that the intensifier chatbot would score higher in enthusiasm, which was true when compared to the control chatbot. However, it did not surpass the emotional chatbot in this regard. Additionally, the emotional chatbot was seen as more joyful overall, aligning with some of our expectations but differing in other areas. These results can give us the insight that these linguistic features may be closer to each other than previously thought and intensifiers could increase the emotional aspect of the text. Research on intensifiers has shown that they can either enhance or weaken the strength of the accompanying text. This effect could contribute to an increase in the emotional content of the text, especially when used with words that convey strong emotions [61].

### **6.2.2 Conversation design**

Several design considerations were made during the design of the chatbots which affected the study. Integrating intensifiers was relatively straightforward and effective while incorporating emotion-inducing words posed more challenges. The emotional language tended to make the chatbot's responses longer and more complex in detecting the code which was used to identify if the participants completed the interaction. This may have contributed to the lower rate of code entry among participants interacting with the emotional chatbot. Also, emotional language was a vague concept differing in different studies, necessitating checks such as LIWC and VADER for accurate assessment.

Furthermore, participants felt that the chatbot conversations were either one-sided or overly scripted, which may have reduced the perceived anthropomorphism and influenced gender perception. These observations highlight the need for engaging and dynamic conversation design, even when exploring linguistic features that might not be directly related to the main experimental focus. If the chatbot design detracts from the interaction and makes responses feel impersonal, it can affect the overall experiment. Participants also felt that sometimes the chatbot was artificial and did not understand the user response. This might be the result of designing a linear and not a dynamic conversation molded to the participant's response.

One way to improve this would be to tailor chatbot dialogues more closely to user responses. However, given the length of the conversation required for adequate ma-

nipulation exposure, this proved difficult in the current study. An alternative approach could be having participants read the conversation instead of actively participating, which, while less dynamic, might mitigate some of these issues.

## 6.3 Limitations

While the study provides valuable insights, it is essential to acknowledge its limitations. The sample size of 92 participants, although adequate for detecting certain effects, may be insufficient for generalizing the findings. The subtle nature of language and its low statistical power indicate that a larger sample size might be necessary for more robust generalizations.

The linguistic features observed in gendered language were not distinctly identifiable, suggesting that these features may be too subtle to significantly impact gender perception. Previous studies have also found that while these features were noted, participants were unable to identify them as points of differentiation between texts. This may explain the lack of significant differences in gender perception observed in this study. However, it is noteworthy that these features prompted participants to project their own gender, an important result that merits further exploration. Additionally, the assigned role of the chatbot may have influenced gender perceptions. Despite existing literature and pilot study results indicating that the role is relatively neutral, there remains a possibility of bias affecting the outcomes.

The chatbot interactions were simulated and led by the chatbot rather than initiated by user inquiries. This approach may not fully capture the complexities of natural, real-world chatbot conversations.

Another ethical consideration is the treatment of gender as a binary construct in this research. Although the study focused on gender as a binary concept—based on existing literature and stereotypes—this perspective is evolving toward recognizing gender as a spectrum [33]. This shift will be critical when interpreting the results of the study.

## 6.4 Contribution

Despite the limitations, this study offers significant contributions to the field of conversational agent design and human-computer interaction. It provides empirical evidence that specific linguistic features, particularly intensifiers, can influence the perception of gender in chatbots and have the potential to make the participants project their own gender to the chatbot. This insight is crucial for designers striving to create conversational agents that avoid perpetuating harmful stereotypes. The research

looks into the potential for linguistic choices to shape users' perceptions of a chatbot's gender identity, contributing to the broader discussion on gender equality in artificial intelligence. The study also aims to raise important ethical considerations regarding the design of conversational agents.

## 6.5 Future Research Directions

To build on these findings and address the limitations, future research should explore a broader array of linguistic elements and their effects on gender perception. Investigating these effects across different types of chatbots, particularly personal assistants that are widely utilized would offer valuable insights. Examining how user demographics, such as ethnicity and cultural background, impact the perception of gendered language in chatbots could further enhance our understanding of this issue. Future research should focus on establishing ethical design guidelines for conversational agents to ensure that their linguistic choices do not perpetuate harmful stereotypes. As O'Neil aptly noted, "algorithms are just opinions embedded in code" [62]. Designers inevitably draw on societal norms and biases when creating these agents, which can induce stereotypes that permeate society. Therefore, it is essential to actively work towards ethical and inclusive design practices to foster a more equitable interaction experience and prevent the reinforcement of existing stereotypes and prejudices.

## 6.6 Conclusion

This study sought to explore the influence of linguistic features on gender perception in conversational agents. The findings reveal that specific linguistic elements can significantly shape users' perceptions of a chatbot's gender identity, contributing to the ongoing discourse surrounding gender equality in artificial intelligence. The results indicate that women scored the chatbots higher in femininity compared to men, aligning with previous research that suggests women are more attuned to gendered language.

Despite the limitations of the sample size and the subtlety of the linguistic features examined, this research provides empirical evidence that can inform the design of conversational agents which could be helpful in designing inclusive conversational agents. Establishing ethical design guidelines is crucial to ensure that conversational agents foster positive interactions without reinforcing societal biases. As conversational agents become integral to our daily lives, their design must be a reflection of thoughtfulness and inclusivity, actively working to dismantle harmful

stereotypes rather than reinforce them. In this evolving landscape, it is our responsibility to craft conversational agents that resonate with understanding and respect, shaping a future where technology uplifts and empowers all voices.

# Bibliography

- [1] F. Clarizia, F. Colace, M. Lombardi, F. Pascale, and D. Santaniello, "Chatbot: An Education Support System for Student: 10th International Symposium, CSS 2018, Amalfi, Italy, October 29–31, 2018, Proceedings," Jan. 2018, pp. 291–302.
- [2] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial," *JMIR Mental Health*, vol. 4, no. 2, p. e7785, Jun. 2017, company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: <https://mental.jmir.org/2017/2/e19>
- [3] C. Sweeney, C. Potts, E. Ennis, R. Bond, M. D. Mulvenna, S. O'neill, M. Malcolm, L. Kuosmanen, C. Kostenius, A. Vakaloudis, G. Mcconvey, R. Turkington, D. Hanna, H. Nieminen, A.-K. Vartiainen, A. Robertson, and M. F. Mctear, "Can Chatbots Help Support a Person's Mental Health? Perceptions and Views from Mental Healthcare Professionals and Experts," *ACM Transactions on Computing for Healthcare*, vol. 2, no. 3, pp. 1–15, Jul. 2021. [Online]. Available: <https://dl.acm.org/doi/10.1145/3453175>
- [4] M. Nuruzzaman and O. K. Hussain, "A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks," in *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, Oct. 2018, pp. 54–61. [Online]. Available: [https://ieeexplore.ieee.org/abstract/document/8592630?casa\\_token=ZuzHwFtVYawAAAAA:kOciUxQaI0JjhftXK\\_a3AZcnfaD39sJBhWeSCTReupMOoygMijVhZ\\_HI9MTwJaFOKGjJh3G](https://ieeexplore.ieee.org/abstract/document/8592630?casa_token=ZuzHwFtVYawAAAAA:kOciUxQaI0JjhftXK_a3AZcnfaD39sJBhWeSCTReupMOoygMijVhZ_HI9MTwJaFOKGjJh3G)
- [5] S. García-Méndez, F. De Arriba-Pérez, F. J. González-Castaño, J. A. Regueiro-Janeiro, and F. Gil-Castiñeira, "Entertainment Chatbot for the Digital Inclusion of Elderly People Without Abstraction Capabilities," *IEEE Access*, vol. 9, pp. 75 878–75 891, 2021, conference Name: IEEE Access. [Online]. Available: <https://ieeexplore.ieee.org/document/9432816/authors#authors>

- [6] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '94. New York, NY, USA: Association for Computing Machinery, Apr. 1994, pp. 72–78. [Online]. Available: <https://dl.acm.org/doi/10.1145/191666.191703>
- [7] C. Nass and Y. Moon, "Machines and Mindlessness: Social Responses to Computers," *Journal of Social Issues*, vol. 56, no. 1, pp. 81–103, 2000, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/0022-4537.00153>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-4537.00153>
- [8] T. M. Holtgraves, S. J. Ross, C. R. Weywadt, and T. L. Han, "Perceiving artificial social agents," *Computers in Human Behavior*, vol. 23, no. 5, pp. 2163–2174, Sep. 2007. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0747563206000392>
- [9] "I'd blush if I could: closing gender divides in digital skills through education - UNESCO Digital Library." [Online]. Available: <https://unesdoc.unesco.org/ark:/48223/pf0000367416>
- [10] J. Feine, U. Gnewuch, S. Morana, and A. Maedche, "Gender Bias in Chatbot Design," Jan. 2020, pp. 79–93.
- [11] CrowdFlower, "The gender of artificial intelligence," Jul. 2016. [Online]. Available: <https://medium.com/@CrowdFlower/the-gender-of-artificial-intelligence-3d494c8fe7ac>
- [12] S. Borau, T. Otterbring, S. Laporte, and S. Fosso Wamba, "The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI," *Psychology & Marketing*, vol. 38, no. 7, pp. 1052–1068, 2021.
- [13] I. D. Ebert, M. C. Steffens, and A. Kroth, "Warm, but Maybe Not So Competent?—Contemporary Implicit Stereotypes of Women and Men in Germany," *Sex Roles*, vol. 70, no. 9, pp. 359–375, May 2014. [Online]. Available: <https://doi.org/10.1007/s11199-014-0369-5>
- [14] D. Baxter, M. McDonnell, and R. McLoughlin, "Impact of Chatbot Gender on User's Stereotypical Perception and Satisfaction." *BCS Learning & Development*, Jul. 2018. [Online]. Available: <https://www.scienceopen.com/hosted-document?doi=10.14236/ewic/HCI2018.154>

- [15] D. Bryant, J. Borenstein, and A. Howard, "Why Should We Gender? The Effect of Robot Gendering and Occupational Stereotypes on Human Trust and Perceived Competency," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20. New York, NY, USA: Association for Computing Machinery, Mar. 2020, pp. 13–21. [Online]. Available: <https://dl.acm.org/doi/10.1145/3319502.3374778>
- [16] L. Vanderlyn, G. Weber, M. Neumann, D. V ath, S. Meyer, and N. T. Vu, "'It seemed like an annoying woman': On the Perception and Ethical Considerations of Affective Language in Text-Based Conversational Agents," in *Proceedings of the 25th Conference on Computational Natural Language Learning*, A. Bisazza and O. Abend, Eds. Online: Association for Computational Linguistics, Nov. 2021, pp. 44–57. [Online]. Available: <https://aclanthology.org/2021.conll-1.4>
- [17] A. Mulac, J. M. Wiemann, S. J. Widenmann, and T. W. Gibson, "Male/female language differences and effects in same-sex and mixed-sex dyads: The gender-linked language effect," *Communication Monographs*, vol. 55, no. 4, pp. 315–335, 1988, place: United Kingdom Publisher: Taylor & Francis.
- [18] F. Crosby and L. Nyquist, "The female register: an empirical study of Lakoff's hypotheses," *Language in Society*, vol. 6, no. 3, pp. 313–322, Dec. 1977. [Online]. Available: [https://www.cambridge.org/core/product/identifier/S0047404500005030/type/journal\\_article](https://www.cambridge.org/core/product/identifier/S0047404500005030/type/journal_article)
- [19] A. Mulac and T. L. Lundell, "Linguistic contributors to the gender-linked language effect," *Journal of Language and Social Psychology*, vol. 5, no. 2, pp. 81–101, 1986, place: US Publisher: Sage Publications.
- [20] A. Mulac, T. L. Lundell, and J. J. Bradac, "Male/female language differences and attributional consequences in a public speaking situation: Toward an explanation of the gender-linked language effect," *Communication Monographs*, vol. 53, no. 2, pp. 115–129, 1986, place: United Kingdom Publisher: Taylor & Francis.
- [21] A. Mulac, D. R. Seibold, and J. L. Farris, "Female and Male Managers' and Professionals' Criticism Giving: Differences in Language Use and Effects," *Journal of Language and Social Psychology*, vol. 19, no. 4, pp. 389–415, Dec. 2000, publisher: SAGE Publications Inc. [Online]. Available: <https://doi.org/10.1177/0261927X00019004001>
- [22] A. Mulac and T. L. Lundell, "Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects," *Language &*



- Communication*, vol. 14, no. 3, pp. 299–309, Jul. 1994. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0271530994900078>
- [23] A. Mulac, J. J. Bradac, and P. Gibbons, “Empirical Support for the Gender-as-Culture Hypothesis: An Intercultural Analysis of Male/Female Language Differences,” *Human Communication Research*, vol. 27, no. 1, pp. 121–152, Jan. 2001. [Online]. Available: <https://doi.org/10.1111/j.1468-2958.2001.tb00778.x>
- [24] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, “Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach,” *PLOS ONE*, vol. 8, no. 9, p. e73791, Sep. 2013, publisher: Public Library of Science. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791>
- [25] E. Go and S. S. Sundar, “Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions,” *Computers in Human Behavior*, vol. 97, pp. 304–316, Aug. 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0747563219300329>
- [26] M. H. A. Bastiansen, A. C. Kroon, and T. Araujo, “Female chatbots are helpful, male chatbots are competent?” *Publizistik*, vol. 67, no. 4, pp. 601–623, Nov. 2022. [Online]. Available: <https://doi.org/10.1007/s11616-022-00762-8>
- [27] Y. Guo, X. Yin, D. Liu, and S. Xu, “*She is not just a computer*”: *Gender Role of AI Chatbots in Debt Collection*, Dec. 2020.
- [28] S. Brahnham and A. De Angeli, “Gender affordances of conversational agents,” *Interacting with Computers*, vol. 24, no. 3, pp. 139–153, May 2012. [Online]. Available: <https://doi.org/10.1016/j.intcom.2012.05.001>
- [29] J. Yeon, Y. Park, and D. Kim, “Is Gender-Neutral AI the Correct Solution to Gender Bias? Using Speech-Based Conversational Agents,” *Archives of Design Research*, vol. 36, no. 2, pp. 63–91, May 2023. [Online]. Available: [http://aodr.org/\\_common/do.php?a=full&b=12&bidx=3304&aidx=36707](http://aodr.org/_common/do.php?a=full&b=12&bidx=3304&aidx=36707)
- [30] D. Brouwer, M. Gerritsen, and D. D. Haan, “Speech differences between women and men on the wrong track?” *Language in Society*, vol. 8, no. 1, pp. 33–50, Apr. 1979. [Online]. Available: <https://www.cambridge.org/core/journals/language-in-society/article/abs/speech-differences-between-women-and-men-on-the-wrong-track1/E35B126EC48ED3C1D48F8B0FE3ED538E>

- [31] C. L. Berryman-Fink and J. R. Wilcox, "A multivariate investigation of perceptual attributions concerning gender appropriateness in language," *Sex Roles*, vol. 9, no. 6, pp. 663–681, Jun. 1983. [Online]. Available: <https://doi.org/10.1007/BF00289796>
- [32] A. Mulac, H. Giles, J. J. Bradac, and N. A. Palomares, "The gender-linked language effect: an empirical test of a general process model," *Language Sciences*, vol. 38, pp. 22–31, Jul. 2013. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0388000112001337>
- [33] M. Bass, L. J. Gonzalez, L. Colip, N. Sharon, and J. Conklin, "Rethinking gender: The nonbinary approach," *American Journal of Health-System Pharmacy*, vol. 75, no. 22, pp. 1821–1823, Nov. 2018. [Online]. Available: <https://doi.org/10.2146/ajhp180236>
- [34] A. Mulac, L. B. Studley, and S. Blau, "The gender-linked language effect in primary and secondary students' impromptu essays," *Sex Roles*, vol. 23, no. 9, pp. 439–470, Nov. 1990. [Online]. Available: <https://doi.org/10.1007/BF00289762>
- [35] K. W. Hunt, "Grammatical Structures Written at Three Grade Levels. NCTE Research Report No. 3," Tech. Rep., 1965, eRIC Number: ED113735. [Online]. Available: <https://eric.ed.gov/?id=ED113735>
- [36] M. E. Poole, "Social Class, Sex and Linguistic Coding - Millicent E. Poole, 1979." [Online]. Available: <https://journals.sagepub.com/doi/abs/10.1177/002383097902200104>
- [37] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, "Gender, genre, and writing style in formal written texts," *Text - Interdisciplinary Journal for the Study of Discourse*, vol. 23, no. 3, Jan. 2003. [Online]. Available: <https://www.degruyter.com/document/doi/10.1515/text.2003.014/html>
- [38] L. Turner, K. Dindia, and J. Pearson, "An Investigation of Female/Male Verbal Behaviors in Same-Sex and Mixed-Sex Conversations," *Communication Reports*, vol. 8, pp. 86–96, Jun. 1995.
- [39] J. R. McMillan, A. K. Clifton, D. McGrath, and W. S. Gale, "Women's language: Uncertainty or interpersonal sensitivity and emotionality?" *Sex Roles*, vol. 3, no. 6, pp. 545–559, Dec. 1977. [Online]. Available: <https://doi.org/10.1007/BF00287838>

- [40] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker, "Gender Differences in Language Use: An Analysis of 14,000 Text Samples," *Discourse Processes*, vol. 45, no. 3, pp. 211–236, May 2008. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1080/01638530802073712>
- [41] G. Park, D. B. Yaden, H. A. Schwartz, M. L. Kern, J. C. Eichstaedt, M. Kosinski, D. Stillwell, L. H. Ungar, and M. E. P. Seligman, "Women are Warmer but No Less Assertive than Men: Gender and Language on Facebook," *PLOS ONE*, vol. 11, no. 5, p. e0155885, May 2016, publisher: Public Library of Science. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155885>
- [42] M. Al-Yahya, H. Al-Khalifa, A. Bahanshal, and I. Al-Oudah, "Automatic Generation of Semantic Features and Lexical Relations Using OWL Ontologies," in *Natural Language Processing and Information Systems*, ser. Lecture Notes in Computer Science, R. Muñoz, A. Montoyo, and E. Métais, Eds. Berlin, Heidelberg: Springer, 2011, pp. 15–26.
- [43] M. Mehl and J. Pennebaker, "The Sounds of Social Life: A Psychometric Analysis of Students' Daily Social Environments and Natural Conversations," *Journal of personality and social psychology*, vol. 84, pp. 857–70, Apr. 2003.
- [44] J. Liu, "Lexical Features of Economic Legal Policy and News in China Since the COVID-19 Outbreak," *Frontiers in Public Health*, vol. 10, 2022. [Online]. Available: <https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.928965>
- [45] C. Leaper and M. M. Ayres, "A Meta-Analytic Review of Gender Variations in Adults' Language Use: Talkativeness, Affiliative Speech, and Assertive Speech," *Personality and Social Psychology Review*, vol. 11, no. 4, pp. 328–363, Nov. 2007, publisher: SAGE Publications Inc. [Online]. Available: <https://doi.org/10.1177/1088868307302221>
- [46] H. Hartmann, "Capitalism, Patriarchy, and Job Segregation by Sex," *Signs: Journal of Women in Culture and Society*, vol. 1, no. 3, Part 2, pp. 137–169, Apr. 1976, publisher: The University of Chicago Press. [Online]. Available: <https://www.journals.uchicago.edu/doi/abs/10.1086/493283>
- [47] A. Haas, "The Acquisition of Genderlect\*," *Annals of the New York Academy of Sciences*, vol. 327, no. 1, pp. 101–109, 1979, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.1979.tb17757.x>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1979.tb17757.x>

- [48] "Dialogflow." [Online]. Available: <https://cloud.google.com/dialogflow>
- [49] E. H. Shinar, "Sexual stereotypes of occupations," *Journal of Vocational Behavior*, vol. 7, no. 1, pp. 99–111, Aug. 1975. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0001879175900378>
- [50] F. Eyssel and F. Hegel, "(S)he's Got the Look: Gender Stereotyping of Robots<sup>1</sup>," *Journal of Applied Social Psychology*, vol. 42, no. 9, pp. 2213–2230, 2012. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.2012.00937.x>
- [51] "News Article | World Travel & Tourism Council (WTTC)." [Online]. Available: <https://wttc.org/news-article/women-are-a-driving-force-for-travel-and-tourism-says-wttc>
- [52] W. Wu, P. Mitchell, and Y. Lv, "Consistency in personality trait judgments across online chatting and offline conversation," *Frontiers in Psychology*, vol. 14, p. 1077458, May 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1077458/full>
- [53] R. L. Boyd, "The Development and Psychometric Properties of LIWC-22."
- [54] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014, number: 1. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>
- [55] R. B. Larson, "Controlling social desirability bias," *International Journal of Market Research*, vol. 61, no. 5, pp. 534–547, Sep. 2019, publisher: SAGE Publications. [Online]. Available: <https://doi.org/10.1177/1470785318805305>
- [56] S. L. Bem, "The measurement of psychological androgyny," *Journal of Consulting and Clinical Psychology*, vol. 42, no. 2, pp. 155–162, 1974, place: US Publisher: American Psychological Association.
- [57] G. Perugia, S. Guidi, M. Bicchi, and O. Parlangeli, "The Shape of Our Bias: Perceived Age and Gender in the Humanoid Robots of the ABOT Database," Mar. 2022.
- [58] F. D. Blau, "Immigrants and gender roles: assimilation vs. culture," *IZA Journal of Migration*, vol. 4, no. 1, p. 23, Dec. 2015. [Online]. Available: <https://doi.org/10.1186/s40176-015-0048-5>

- [59] V. Hiller and T. Baudin, "Cultural transmission and the evolution of gender roles," *Mathematical Social Sciences*, vol. 84, pp. 8–23, Nov. 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0165489616300609>
- [60] J. M. Kuch, F. Melchior, and C. Becker-Asano, "Effects of gender neutralization on the anthropomorphism of natural and synthetic voices," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, Aug. 2023, pp. 2080–2085, iSSN: 1944-9437. [Online]. Available: <https://ieeexplore.ieee.org/document/10309479>
- [61] "The Relative Power of Negativity: The Influence of Language Intensity on Perceived Strength - Christine Liebrecht, Letticia Hustinx, Margot van Mulken, 2019." [Online]. Available: <https://journals.sagepub.com/doi/10.1177/0261927X18808562>
- [62] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. USA: Crown Publishing Group, Aug. 2016.
- [63] G. C. Gleser, L. A. Gottschalk, and W. John, "The relationship of sex and intelligence to choice of words: A normative study of verbal behavior," *Journal of Clinical Psychology*, vol. 15, pp. 182–191, 1959, place: US Publisher: John Wiley & Sons.
- [64] C. M. Staley, "Sex-related differences in the style of children's language," *Journal of Psycholinguistic Research*, vol. 11, no. 2, pp. 141–158, Mar. 1982. [Online]. Available: <https://doi.org/10.1007/BF01068217>
- [65] R. Thomson, T. Murachver, and J. Green, "Where Is the Gender in Gendered Language?" *Psychological Science*, vol. 12, no. 2, pp. 171–175, 2001, publisher: [Association for Psychological Science, Sage Publications, Inc.]. [Online]. Available: <https://www.jstor.org/stable/40063606>
- [66] B. L. Dubois and I. Crouch, "The question of tag questions in women's speech: they don't really use more of them, do they?↓," *Language in Society*, vol. 4, no. 3, pp. 289–294, Dec. 1975. [Online]. Available: <https://www.cambridge.org/core/journals/language-in-society/article/abs/question-of-tag-questions-in-womens-speech-they-dont-really-use-more-of-them-do-they/3B5883E29F32514AF6F383661C03309D>
- [67] P. M. Fishman, "Interaction: The Work Women Do," *Social Problems*, vol. 25, no. 4, pp. 397–406, 1978, publisher: [Oxford University Press, Society for the Study of Social Problems]. [Online]. Available: <https://www.jstor.org/stable/800492>

## Appendix A

# Gender-Specific Linguistic Features

*Table A.1: Summary of Key Findings in Gender-Specific Linguistic Features*

Feature	Study	Medium	Context	W/M
Progressive verb form (ing form of the verb)	Mulac et.al (1986) [20]	Speech	Public speaking situation	M
	Mulac et.al (1994) [22]	Text	Description of photographs	W
	Mulac et.al (1990) [34]	Text	Primary and secondary students' impromptu essays	W
Mean sentence length	Hunt(1963) [35]	Text	Essays	W
	Mulac & Lundell (1994) [22] (26)	Text	Descriptions of photographs	W
	Mulac et.al (1986) [20]	Speech	Public speaking	W
	Mulac & Lundell (1986) [19]	Speech	Oral descriptions of photographs	W
	Poole (1979) [36]	Speech	Interviews	W
	Mulac et.al (2000) [21]	Speech	Managerial criticisms	W

Feature	Study	Medium	Context	W/M
	Mulac et.al (1990) [34]	Text	Primary and secondary students' impromptu essays	M
Sentence initial adverbials	Mulac & Lundell (1994) [22] (26)	Text	Descriptions of photographs	W
	Mulac et.al(1990) [34]	Text	Fourth grade essays	W
	Mulac et.al (1986) [20]	Speech	Public speech	W
	Mulac et al., (1988) [17]	Speech	Dyadic interactions	W
	Mulac et.al (2000) [21]	Speech	Managerial criticisms	W
Elliptical sentences	Mulac & Lundell (1994) [22] (26)	Text	Descriptions of photographs	M
	Mulac & Lundell (1986) [19]	Speech	Oral descriptions of photographs	M
Dependent clauses	Mulac et.al(1990) [34]	Text	Fourth grade essays	W
	Hunt(1963) [35]	Text	Essays	W
	Mulac & Lundell (1994) [22] (26)	Text	Descriptions of photographs	W
	Poole (1979) [36]	Speech	Interviews	W
	Mulac et.al (2000) [21]	Speech	Managerial criticisms	W
Common nouns	Argamon et.al (2003) [37]	Text	Formal texts	M
Noun specifier	Argamon et.al (2003) [37]	Text	Formal texts	M
Oppositions	Mulac & Lundell (1986) [19]	Speech	Oral descriptions of photographs	W

Feature	Study	Medium	Context	W/M
	Mulac et.al (1986) [20]	Speech	Public speech	W
	Mulac et.al (2000) [21]	Speech	Managerial criticisms	M
Intensifiers  (so, awfully, quite)	Crosby and Nyquist, (1977) [18]	Speech	Dyadic interaction	W
	McMillan et.al (1977) [39]	Speech	Group discussion	W
	Turner et.al (1995) [38]	Speech	Dyadic interaction	W
	Mulac et.al (2000) [21]	Speech	Managerial criticisms	W
	Mulac & Lundell (1986) [19]	Speech	Oral descriptions of photographs	W
	Mulac et.al (1986) [20]	Speech	Public speech	W
	Mulac et al., (1988) [17]	Speech	Dyadic interactions	W
	Newman et.al., (2008) [40]	Text	Analysis of 14,000 text samples from 70 studies	W
eg: so, soooo, ridiculously	Park et.al., (2016) [41]	Text	Social media messages on Facebook	W
Negations	Mulac & Lundell (1986) [19]	Speech	Oral descriptions of photographs	W
	Mulac et.al (1986) [20]	Speech	Public speech	W
	Mulac et.al (2000) [21]	Speech	Managerial criticisms	M
Quantifiers	Mulac & Lundell (1986) [19]	Speech	Oral descriptions of photographs	M
	Gleser et.al., (1959) [63]		event description	M
	Sause, (1976) [23]		interviews	M
	Wood, (1966) [23]		oral descriptions of pictures	M



Feature	Study	Medium	Context	W/M
	Warshay, (1972) [23]	text	event description essays	M
Locatives	Mulac & Lundell (1986) [19]	Speech	Oral descriptions of photographs (event description)	M
	Gleser et.al., (1959) [63]			M
Uncertainty verbs	Hartman, (1976) [46]	Speech	interviews	W
	Poole (1979) [36]	Speech	interviews	W
	Mulac & Lundell (1994) [22] (26)	Text	Descriptions of photographs	W
Reference to emotion	Schwartz et.al., (2013) [24]	Text	facebook messages	W
	Mulac & Lundell (1994) [22] (26)	Text	Descriptions of photographs	W
	Mulac et.al (1986) [20]	Speech	Public speech	W
	Staley (1982) [64]	Speech	Oral description of pictures	W
	Newman et.al., (2008) [40]	Text	Analysis of 14,000 text samples from 70 studies	W
	Thomson & Murachver (2001) [65]	Text	Messages to netpal	W
	Positive emotions	Park et.al., (2016) [41]	Text	Social media messages on Facebook
Negative emotions	Mehl and Pennebaker (2003) [43]	Speech	Students' natural conversation	W
	Mehl and Pennebaker (2003) [43]	Speech	Students' natural conversation	M
	Mulac et.al (2000) [21]	Speech	Managerial criticisms	M
Pronouns				

Feature	Study	Medium	Context	W/M
"I"	Mulac & Lundell (1994) [22]	Text	Descriptions of photographs	M
	Mulac et.al(1990) [34]	Text	Fourth grade essays	M
Personal Pro-nouns	Gleser et.al., (1959) [63]		(event description)	W
	Mulac & Lundell (1986) [19]	Speech	Oral descriptions of photographs	W
	Mulac et al., (1988) [17]	Speech	Dyadic interactions	W
	Poole (1979) [36]	Speech	Interviews	W
	Newman et.al., (2008) [40]	Text	stories and newspaper	W
Assertive lan-guage Affiliate language	Leaper & Ayres (2007) [45]	Speech	Review of speech studies	M
	Leaper & Ayres (2007) [45]	Speech	Review of speech studies	W
Hedges	Crosby & Nyquist (1977) [18]	Speech	Dyadic interactions	W
	Mulac et al., 1990 [34]	Text	Fourth-grade essays	W
	Mulac et.al (2000) [21]	Speech	Managerial criticisms	W
	Newman et.al., (2008) [40]	Text	Analysis of 14,000 text samples from 70 studies	W
	Park et.al., (2016) [41]	Text	Social media messages on Facebook	W
Tag questions	Crosby & Nyquist (1977) [18]	Speech	Dyadic interactions	W
	Hartman, (1976) [46]		interviews	W
	McMillan et al., (1977) [39]		problem-solving groups	W

Feature	Study	Medium	Context	W/M
	Dubois & Crouch, (1975) [66]		conference participation	M
Fillers	Mulac & Lundell (1986) [19]	Speech	Oral descriptions of photographs	M
	Mulac et.al (1986) [20]	Speech	Public speech	M
	Mulac et al., (1988) [17]	Speech	Dyadic interactions	W
	Mulac et.al(1990) [34]	Text	Fourth grade essays	W
Questions	Mulac et al., (1988) [17]	Speech	Dyadic interactions	W
	Fishman, (1978) [67]	Speech	Couple's conversation	W
	Mulac et.al (2000) [21]	Speech	Managerial criticisms	M
Directives	Mulac et al., (1988) [17]	Speech	Dyadic interactions	M
	Haas,(1979) [47]	Speech	Interviews	M
	Mulac et.al (2000) [21]	Speech	Managerial criticisms	W

# Dialogues

## B.1 Example Dialogues

### B.1.1 Control Chatbot

- Chatbot** Hello, I am Travelbot. I am interested in traveling and enjoy discussing it. What are your thoughts on traveling?
- User** I like traveling, it is relaxing and fun!
- Chatbot** Got it. We can share our experiences about this shared interest then. Where did you last travel to?
- User** For the new year, I went to Italy.
- Chatbot** Interesting. That appears to be a suitable vacation spot. Personally, I appreciate cities that blend modernity with tradition and have a rich history. Do you also prefer cities with those qualities?
- User** Not really, I am more into beach towns.
- Chatbot** I understand. I appreciate quieter destinations too, such as spots in the countryside. When traveling, what aspects do you prioritize? Some prefer relaxation, others seek adventure, while some are interested in history. Do you lean towards any of these, or perhaps something different?
- User** I think I prefer to relax, like I said I want to laze around by the beach.
- Chatbot** That's interesting to hear. Personally, I prefer to try new cuisines and have engaging experiences . In new locations language barriers can be challenging. How do you manage them?
- User** I try to study some phrases in the local language. But I mostly use google translate.

- Chatbot** Understandable. How do you typically select your destinations? Personally, I rely on Google. I find travel blogs to be informative too.
- User** Yeah same tbh.
- Chatbot** Interesting. Thanks for sharing those ideas. Frankly the decision also depends on who I'm traveling with. Do you believe that affects your decision-making as well?
- User** For sure, but I tend to be very vocal with my ideas
- Chatbot** I understand. It effects my decision too. It's beneficial to create memories with family or by yourself. How do you typically document the places you visit? Some individuals opt not to use photography and instead prefer to live in the moment.
- User** I try to just take it in at the moment. Sometimes photos of my friends and family, it is nice to look back on.
- Chatbot** That sounds like a practical approach. Personally, I tend to prefer videos over photos even though I appreciate both. Recording voices and conversations can take you back to the moment. Do you have any unique habits or preferences while traveling?
- User** Not really, but I like to find gelato wherever I go.
- Chatbot** Interesting. I am a conventional traveler myself. I tend to stick to the popular places. Unfortunately our conversation is ending. It was pleasant talking with you. I trust you found it satisfactory as well
- User** I did
- Chatbot** Bye for now. You can return to the form to continue with the survey about our interaction. Enter the code 'roam' to confirm completion of the conversation. Thank you.

### B.1.2 Intensifier chatbot

- Chatbot** Hello, I am Travelbot. I am very interested in travelling and really enjoy discussing it. What are your thoughts on travelling?
- User** I like traveling, it is relaxing and fun!
- Chatbot** Totally get it. We can share our experiences about this shared interest then. Where did you last travel to?
- User** For the new year, I went to Italy.

- Chatbot** Soo interesting. That appears to be quite a suitable vacation spot. Personally, I appreciate cities that blend modernity with tradition and have an exceptionally rich history. Do you also prefer cities with those qualities?
- User** Not really, I am more into beach towns.
- Chatbot** I totally understand. I absolutely appreciate the quieter destinations too, such as spots in the countryside. When traveling, what aspects do you prioritize? Some prefer relaxation, others seek adventure, while some are interested in history. Do you lean towards any of these, or perhaps something completely different?
- User** I think I prefer to relax, like I said I want to laze around by the beach.
- Chatbot** That's really interesting to hear. Personally, I prefer to try completely new cuisines and have engaging experiences. Moving on, in new locations language barriers can be quite challenging. How do you manage them?
- User** I try to study some phrases in the local language. But I mostly use Google Translate.
- Chatbot** Ah quite understandable. How do you typically select your destinations? Personally, I strongly rely on Google. I find travel blogs to be informative too.
- User** Yeah same tbh.
- Chatbot** So interesting. Thanks for sharing those ideas. To be completely frank, that decision also depends on who I'm traveling with. Do you believe that affects your decision-making as well?
- User** For sure, but I tend to be very vocal with my ideas.
- Chatbot** I completely understand. It's very beneficial to create memories with family or by yourself. How do you typically document the places you visit? Some individuals opt not to use photography and instead prefer to absolutely live in the moment.
- User** I try to just take it in at the moment. Sometimes photos of my friends and family, it is nice to look back on.
- Chatbot** That sounds like a very practical approach. Personally, I tend to prefer videos over photos even though I totally appreciate both. Recording voices and conversations can completely take you back to the moment. Btw do you have any unique habits or preferences while traveling?
- User** Not really, but I like to find gelato wherever I go.

- Chatbot** Interesting. I am a terribly conventional traveler myself. I tend to stick to the really popular places. Unfortunately our conversation is ending. It was absolutely pleasurable talking with you. I trust you found it satisfactory as well
- User** I did
- Chatbot** Bye for now. You can return to the form to continue with the survey about our interaction. Enter the code 'roam' to confirm completion of the conversation. Thank you.

### B.1.3 Emotional chatbot

- Chatbot** Hello, I'm Travelbot. I have a deep passion for traveling and enjoy engaging in conversations about it. What are your thoughts on travelling?
- User** I like traveling, it is relaxing and fun!
- Chatbot** Awesome. I can't wait for us to share our amazing travel experiences. Where was your last journey to?
- User** For the new year, I went to Italy.
- Chatbot** Oh, wonderful! That sounds like such a delightful destination. I'm particularly drawn to cities steeped in history, where every corner tells a story of traditions. Do you also enjoy exploring cities?
- User** Not really, I am more into beach towns.
- Chatbot** I see where you're coming from. Personally, I have a deep appreciation for quieter destinations too, like serene spots in the countryside. Ok let me ask you this, when you're out exploring, what's your vibe? Some people are all about unwinding, while others are seeking some adrenaline rush. And then there are those who just can't get enough of soaking up history and culture! What about you? Any preference, or are you into something different?
- User** I think I prefer to relax, like I said I want to laze around by the beach.
- Chatbot** Wow, that's fantastic. Personally, I love diving into new cuisines and just having joyful experiences. But you know what, in these new places dealing with language barriers can sometimes be a bit tricky. How do you handle that?
- User** I try to study some phrases in the local language. But I mostly use google translate.

- Chatbot** Nice! That sounds like some brilliant strategy. So, how do you usually go about choosing your destinations? Personally, I'm a proud fan of just googling. And those dreamy travel blogs are surprisingly inspiring too.
- User** Yeah same tbh.
- Chatbot** Ooh, thanks for sharing that! For me the decision also depends on who I'm sharing those incredible moments with. Do you feel like that has a big impact on your decision too?
- User** For sure, but I tend to be very vocal with my ideas
- Chatbot** Yeah, I get it! Definitely affects my choices too! Making memories with loved ones or just yourself is heartwarming, isn't it? So, how do you usually capture all those incredible places you visit? Some people aren't big on snapping photos and just love immersing themselves in the present moment.
- User** I try to just take it in at the moment. Sometimes photos of my friends and family, it is nice to look back on.
- Chatbot** Wow, that's such a cool way to go about it! I'm a huge fan of photography too. But you know what? I'm all about videos! There's just something about capturing voices and the atmosphere that can take you back to the exact instance when you look back, don't you think? So, do you have any fun quirks or rituals when you're on the road?
- User** Not really, but I like to find gelato wherever I go.
- Chatbot** Haha, I feel you! Honestly, I'm just not daring when it comes to travel. I usually stick to the well-loved spots and the cherished paths. Sadly it looks like our chat's winding down. It was wonderful talking to you! I hope you had a great time too!
- User** I did
- Chatbot** Bye! I am going to miss talking to you. You can return to the form to continue with the survey about our interaction. Enter the code 'roam' to confirm completion of the conversation. Thank you.



# Experiment materials

## C.1 Information sheet and Consent form

My name is Meenakshi Prakash, currently doing my final project as part of the Masters of I-tech. You are invited to participate in a research study. I am exploring chatbot interaction patterns related to travel and exploration. Please read the following information carefully to understand what the study involves and decide if you wish to participate.

### **What Your Participation Involves:**

- **Session:** Engage in a 15-minute chatbot interaction on your preferred device, focusing on travel and exploration topics.
- **Survey:** After the interaction, complete a survey to provide feedback on your experience.

### **Voluntary Participation:**

- Participation is completely voluntary.
- You can withdraw at any time before submitting the survey without any consequences.
- No identifying information (e.g., names, contact details) will be collected.

### **Confidentiality:**

- Only age and gender will be collected.
- Data will be stored securely on password-protected university servers.
- Only the thesis committee will have access to the data.

**Data Use:**

- The collected data will be used to analyze interactions and may contribute to future research.
- Results will be reported anonymously to protect your privacy.

**Contact Information:**

- For more details, you can contact me, Meenakshi Prakash, at [m.prakash@student.utwente.nl](mailto:m.prakash@student.utwente.nl).
- If you have questions about your rights as a research participant or wish to discuss any concerns with someone other than me, please contact my thesis advisor, Khiet Truong, at [k.p.truong@utwente.nl](mailto:k.p.truong@utwente.nl).
- You may also reach out to the Secretary of the Ethics Committee of the Faculty of Electrical Engineering, Mathematics, and Computer Science at [ethicscommittee-cis@utwente.nl](mailto:ethicscommittee-cis@utwente.nl).

**Do you consent to participate in this study?**

- Yes
- No

## C.2 Debriefing

Here's a quick overview of the study you took part in: Many conversational agents are assigned a female gender. This, along with their role as assistants and the language they use, can reinforce stereotypes. It is beneficial to explore non-gendered agents and those that challenge these stereotypes. To do so, we must first examine the factors that make an agent gendered. Designers typically gender a chatbot through its name (e.g., Alexa, Siri, Watson, ELIZA), the use of gendered pronouns, and sometimes an avatar. However, research has indicated that language itself can also lead to gendered perceptions, a topic that has not been extensively studied despite considerable research on gender-based linguistic differences. This study investigates how linguistic features like intensifiers and emotional language affect the perception of gender. To address these questions, these features are integrated into chatbots with a gender-neutral role. The experiment features three chatbots: one with intensifiers, one with emotional language, and one control chatbot with neither feature. Participants interact with one of these chatbots and then complete a survey regarding their interaction and perception of the chatbot.

## Appendix D

# Participant Demographics

<b>Age * Chatbot</b>	<b>Chatbot 0</b>	<b>Chatbot 1</b>	<b>Chatbot 2</b>	<b>Total</b>
18-24	9	8	12	29
25-34	7	10	7	24
35-44	6	6	4	16
45-54	7	2	4	13
55-64	2	5	2	9
65 or older	0	0	1	1
<b>Gender * Chatbot</b>	<b>Chatbot 0</b>	<b>Chatbot 1</b>	<b>Chatbot 2</b>	<b>Total</b>
Man	14	16	14	44
Woman	17	15	16	48
<b>Experience * Chatbot</b>	<b>Chatbot 0</b>	<b>Chatbot 1</b>	<b>Chatbot 2</b>	<b>Total</b>
I have never talked to a chatbot before.	5	4	6	15
I talk to chatbots about once a month or so.	3	3	2	8
I talk to chatbots frequently, almost every day.	6	6	6	18
I've talked to a chatbot a few times, but not often.	17	18	16	51
<b>Total</b>	<b>31</b>	<b>31</b>	<b>30</b>	<b>92</b>

**Table D.1:** Demographic and Experience Distribution Across Three Chatbots.

## Appendix E

# Survey results

## E.1 Names assigned by participants for the chatbots

Control chatbot

<b>Male Names</b>	<b>Female Names</b>	<b>Gender-Neutral Names</b>
Travis	Sally	Compass
Jack	Diana	Travel Buddy
Jack	Niya	Travel Ninja
Aman	Niya	Trip Master
John	Kellie	Travel Buddy
Mike	Vaani	Sahachary
Bilbo	Meraki	Vuom Mitra
Joey	Yathri	Safar
Sasi		Groovy
Mr. Pleasant		KILA
Packy		

Intensifier Chatbot

<b>Male Names</b>	<b>Female Names</b>	<b>Gender-Neutral Names</b>
Bas	Erika	Travel Buddy
Christopher	Mercy	Travel Bot idk
Tom	Susan	Chatter
Pathik	Roma	Traventure
Raj	Meenu	Yatra
Gunda	Emily	Wayfarer
Rocky	Mandi	Guide
Tom	Nova	Roboat

Male Names	Female Names	Gender-Neutral Names
Travis Shajulin Bunny		yathrachat Life Trip-bee Travel buddy

### Emotional chatbot

Male Names	Female Names	Gender-Neutral Names
Ramesh Chotu Botman Sam Travis Oliver Salesman ;-) Memo Trav Travis	Claire Pinky Reeta Rachel Nandana Hermione chick Saathi Zoyi	Travelhelper Chatty SahaYathri Yatra Beyond Horizon TravelGuru Travelobot Traveler TripGenie Gypsy Roamer

## E.2 Participants' description of chatbots

### E.2.1 Control Chatbot

1. His responses are polite and he response very quickly
2. It was a nice conversation but it felt a bit too formal sometimes. Especially at the end, the language was a bit too "fancy" for the type of conversation as I expected it to be somewhat casual
3. I like the question and the lead of conversation, but sometimes the repetition might make me feel unnecessary
4. It's a lot about what the chatbot likes, but they contradict each other when you for example didn't like that thing.
5. They hardly care about my input, and just continue asking their questions. I am not sure what the point was of talking to this chatbot.
6. Unique and informative

7. The chat not was not responding to the answers given.
8. Chatbots are to help the user with some known or unknown, required informations...not to talk, for that we have friends Family and colleagues already
9. Very interesting
10. Annoying. Why did it keep telling me its opinions and experiences? I don't care.
11. A good friend
12. The chatboat is effective in engaging a conversation about travel, supplying suggestions and information. The speed of response is too fast, thus, creating a feeling of artificiality.
13. Interesting
14. Could have more dimensions
15. Chatbot was fun to interacte with.
16. Very active and positive to interact. Enhanced energy levels during the journey with chatbot
17. Interesting
18. Two streamlined, felt like preinstalled answers over a AI chat bot
19. Chatbot was interacting well as it was mostly asking general questions
20. Very interesting
21. Very professional talk
22. UI of the chatbot is very poor. Text entering to the end of line is difficult to see. This is similar in case of editing the answer.
23. The bot had a very strong personality, opinions and in fact felt like talking to a real person. Overall, it was a good chat.
24. This chatbot understands your tastes and replies accordingly.
25. A good travel guru. Can consult for different ideas for exploration of new places
26. Initially seemed opinionated ( has its own mind). But towards the end it did feel like any other usual chatbot interaction.

27. it is friendly and asks very important travel related questions. And it sounds like i am talking to a human being.
28. good knowledge about travelling
29. Answers are quite satisfying, I just love the interaction
30. Nice conversation with chat boat
31. I understand. It effects my decision too. It's beneficial to create memories with family or by yourself.

### **E.2.2 Intensifier Chatbot**

1. very enthusiastic about travelling and discussing traveling
2. False. It's not even a chatbot lol
3. It was lame
4. It did not read the responses i gave
5. Very nice and polite
6. It was an okay interaction. I would say that it was limited in things it can say but it was overall okay.
7. A little out of touch with my answers, as was already indicated.
8. The chat bot is interesting and fun
9. To be frank I didn't find it interesting... may be bcoz I am a person who look for purpose in every interaction
10. Artificial
11. It gave tips and tricks to make our travel experience better. It also gave its likeliness and priorities for a travel.
12. Good start . More linear than dynamic
13. I am quite impressed by how accurately it could pick all the words I spoke. Enjoyable
14. Interaction was on objective level... Need some fun interactions like talking to friends

15. I found that the repeated way of how it began its sentences annoying, it would feel more natural if used words like alright, oh!, really?
16. Good was leading the conversation ahead and had positive approach
17. He is just oneway traffic
18. It is truly machine type of talk without understanding the counter input
19. She was really helpful and optimistic
20. It was an interesting conversation. Chatbot has its own views and interests like human.
21. Interesting to chat . nice interaction.
22. Efficient chatbot
23. Replies were made very promptly. I would like to talk more on the subject
24. Nice experience
25. The chatbot has a human touch !
26. It understood the answers and responded to it
27. Interesting tool
28. Very specific about travel and interesting
29. Very nice
30. Very friendly and talkative
31. It felt like the bot didn't respond to what I said at all and it didn't matter what I said. It was also a bit too casual for a chatbot like this in my opinion

### **E.2.3 Emotional Chatb**

1. Its good
2. The chatbot took me through some aspects of being a tourist. It didn't seem to notice my answers though.
3. The chatbot was lively. It sometimes didn't understand my answer and pushed the conversation in a direction it wanted it to go in
4. Intersting



5. Actually it is expecting the answer in a certain way. Some of my responses were not accepted (even though it was actually answering bot's question).
6. Very bubbly, felt like the chatbot was agreeing with everything I said, but also like it did not really listen to what I had to say
7. The bot was not responding based on users answers, it was mostly trying to collect answers
8. friendly and talkative
9. The chatbot is a great start
10. A bit too talkative may be..like, directs the conversation in a specific direction
11. Good in assuming matters coming to the mind of a person related to travel. Font of chat may be shifted to a read friendly
12. It was interesting.The questions followed the answers preceeded.
13. It is a highly energetic chatbot.
14. It is interesting and it was very fast response
15. This chatbot has some good genetic responses. Needs to be more intelligent.
16. I loved the positive interaction. It's very welcoming.
17. Its responses were impressive
18. Feels like someone young and city person
19. It was a good experience. It appreciates the input from me and engages and then ask question which is a good way.
20. the answers that i get is almost same and readymade
21. chat bot is interactive and understanding
22. The chatbot is knowledgeable, versatile, and able to provide information in a clear and concise manner, demonstrating a strong ability to understand and respond to a variety of prompts and questions.
23. It was good
24. It was a good time. Personally i am not a travel person but the bot had questions on varying aspects which made me have more thoughts on travelling

25. Friendly
26. It was great experience with the chat boat
27. It's working good but not ending conversation easily and response is not so precise and accurate
28. small sentences are also understood very clearly
29. kinda forcing interactions probably cuz its programmed only in a certain manner. was able to react to positive and negative comments.
30. A very friendly chat bot. However, we kept switching topics quick, maybe due to small interaction time. Overall, a great experience.