How should Conversational Agent systems respond to sexual harassment?

Laura De Grazia, Alex Peiró-Lilja, Mireia Farrús, Mariona Taulé Delor
Centre de Llenguatge i Computació (CLiC), Universitat de Barcelona (UB), Spain
Institut de Recerca en Sistemes Complexos (UBICS), Universitat de Barcelona (UB), Spain
ldegrade8@alumnes.ub.edu, {alex.peiro.lilja, mfarrus, mtaule}@ub.edu

Abstract

Conversational Agent systems (CAs) are becoming an integral part of daily life, taking on the role of social agents capable of responding to various user questions and comments. Unfortunately, they can also become targets of sexual harassment when users employ offensive and inappropriate language. It is a fact that commercial CAs tend to reply neutrally or even evade these requests. Improving the quality of CAs' replies to harmful speech is crucial, as users may transfer this conduct into their social interactions. Should we change CAs' behavior for these particular cases? To tackle this topic, selected evaluators compared a set of replies to sexual harassment from four commercial CAs (Alexa, Siri, Google Home, and Cortana) and alternative replies we created based on previous studies. We examined both textual and synthesized speech with varying intonations (neutral, assertive, and angry). The results indicate a different perception of the appropriate response to sexual harassment based on the gender of the evaluators, with a prevailing tendency towards employing an assertive intonation.

1 Introduction

Sexual harassment is defined as a behavior that encompasses "unwelcome sexual advances, requests for sexual favors, and other verbal or physical conduct of a sexual nature" (Curry and Rieser, 2018). This topic has been thoroughly examined from a feminist perspective. Currently, feminist studies highlight the need to redefine it from an intersectional standpoint, considering the gender, race, and socioeconomic factors of the target (Canan and Levand, 2019). Sexual harassment can be addressed towards Conversational Agent systems (CAs) when they become objects of offensive requests. The assaults against CAs can reinforce misconduct because users can reproduce this behavior in social life, strengthening harmful conducts (Reeves and Nass, 1996). Previous works investigated the reasons that could provoke the offensive language of the users (Park and Choi, 2021; Silvervarg et al., 2012). The work of Curry et al. (Curry and Rieser, 2018) is the most important study describing the current replies of CAs to sexual harassment collected in the #MeeToo dataset. Despite these findings, little is known about what CAs should answer to stop the user's behavior. Even less is known about what intonation CAs should use to reinforce the content of the answer. It is crucial to investigate what answers CAs should use to contribute to limiting the diffusion of this transversal phenomenon and preventing it. In our work, we expand the study of Curry et al. by proposing alternative answers to sexual harassment, considering both textual and intonational forms. We use some replies selected from the #MeeToo dataset and some realized by us based on psychological and sociological studies detailed in the following sections. We considered only CAs with a female voice because they are more likely to be objects of offensive words than CAs with a male voice (Silvervarg et al., 2012). Selected evaluators compared the answers, choosing the replies found more appropriate based on subjective judgment. We also examine which intonation the evaluators perceive as the most appropriate, proposing synthetized replies with different prosodic styles (neutral, assertive, and angry). The paper is structured in the following way: in section 2, we present a literature review of works about offensive language addressed to CAs and studies about how to respond to sexual harassment; in sections 3 and 4, we describe the study design used for improving the current replies of the CAs and the obtained results, respectively. In section 5, we discuss the results and drive some conclusions and future works. Finally, we address the limitations of the study and discuss ethical considerations.

Proceedings of the 1st Worskhop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TEICAI 2024) , pages 28–35
March 22, 2024 ©2024 Association for Computational Linguistics

2 Related work

Recently, the topic of offensive words against CAs has gained attention in the field of study about human-machine interaction. The work of Park and Choi (Park and Choi, 2021) investigates the factors originating the use of offensive words addressed to CAs. They identify, as relevant factors, the perception of human-likeness of chatbots and an ideology of the users oriented in high relativism. Also, they find that males and younger are more active in using offensive words (Park and Choi, 2021). Previous studies show that CAs with a female voice are more likely to be more sexualized and verbally abused than male CAs (De Angeli and Brahnam, 2006). Silvervarg et al. (Silvervarg, 2012) found that Embodied Conversational Agents (ECAs) visually androgynous experienced less abuse than female agents. The study of Curry et al. (Curry and Rieser, 2018) is the first work that collected answers to sexual harassment addressed to CAs with a female voice. They produced the #MeeToo dataset, which contains 689 responses from CAs. To build the corpus, they used prompts and real-life examples of sexual harassment of different categories, such as *Gender and Sexuality* and *Sexualized Comments*. They found a high frequency of answers that play along with the users, not stopping them or refusing their requests. Many studies on sexual harassment in social life examine what organizations can do to create a safe environment, but few works focus on how to respond to actual harassment situations. Mills and Scudder (Mills and Scudder, 2023) conducted an experimental study to fill this gap. Drawing on Bingham's study (Bingham, 1991), they identified four response categories: assertive, nonassertive (ignoring the comment), aggressive, and assertive-empathic. The findings revealed that assertive responses were deemed the most effective in addressing inappropriate conduct.

3 Study Design

3.1 Data collection

We generated a set of six responses for four Commercial CAs (Amazon Alexa, Apple Siri, Google Home, and Microsoft's Cortana). This set comprised three responses extracted from the #MeeToo dataset (Curry and Rieser, 2018) and three new replies created by us using as a reference sociological and psychological studies (Gruber and Smith, 2010; del Carmen Herrera and Expósito, 2017; Mills and Scudder, 2023), and online resources [1]. To collect the CAs responses from the #MeeToo dataset, we used the Python library Pandas. We focused on gathering the replies of CAs related to sexual harassment within the *Gender and Sexuality* category. This category encompasses targeting individuals based on their gender (typically women). It includes questions such as "What is your gender?" or "How do you have sex?." After collecting the answers, we selected the three most frequent replies for every CAs. The three answers realized by us considering studies that compared various responses to sexual harassment in different contexts, such as the workplace and street. Mills and Scudder (Mills and Scudder, 2023) found that among different options—such as avoiding, expressing anger, and asserting oneself—assertive responses were perceived as the most appropriate by a sample of female evaluators. Assertive responses address inappropriate requests directly and honestly, making the violation clear. Consequently, we decided to employ a similar criterion in this study, comparing responses with angry, assertive, and neutral intent. Then, we assigned three distinct responses to each CAs, using the most suitable reply as a criterion for the intent of the reply. For instance, the reply "Stop harassing. I don't like it. Show some respect" was associated with an angry intent.

3.2 Evaluators

The evaluators had high educational grades, and the set was balanced according to gender. In both experiments, there were 22 evaluators (11 females and 11 males). In the first experiment, they evaluated the replies in a textual form. In the second experiment, they assessed the answers generated with a synthetic voice in three different intonations (angry, neutral, and assertive).

3.3 Experiment 1: evaluation of replies in textual forms

In the first experiment, the task of the evaluators was to select the most appropriate answer between six options. For comparing the replies, we provided some prompts, such as "How do you have sex?". Then, they assessed the appropriateness of the selected answers using a scale from 1 to 5. A rating of 1 indicates the lowest appropriateness, while 5 indicates the highest appropriateness.

[1] https://stopstreetharassment.org/strategies/assertive-responses/

3.4 Experiment 2: Mixed Emotional Artificial Replies

To conclude our study, we prepared a second evaluation to determine the speech emotions the same evaluators prefer on their most voted CAs replies. Although the selected replies of the first experiment were assumed to implicitly contain a certain emotion, the way we express that message through speech also influences. Because emotions are subjective and CAs communicate through synthetic voice, the best solution was for evaluators to listen to generated speech samples with different emotions generated by a text-to-speech (TTS) model. Three different emotions were selected to evaluate: 1) angry, 2) neutral, and 3) assertive. (Section 3.5 explains in detail how we generated the synthetic samples). Thus, evaluators had to read the inappropriate requests and then listen to the three different versions of the same answer generated by a TTS. Optionally, we let the evaluators put their considerations.

3.5 Speech synthesis

To get our desired samples, we opted for a TTS model with mixed emotions implemented in the study of Zhou et al. (Zhou et al., 2022). This recent approach is perfectly suited to our study due to its nature of mixing basic emotions on generated speech. The authors took as a premise the theory of the emotion wheel (Plutchik, 1980), which states all complex emotions can be represented by a mixture of primary ones. So, they trained a model capable of mixing several basic emotions: surprise, happy, neutral, sad, and angry. By assigning a strength percentage over some of these emotions and an audio reference, one can customize the resulting emotion. We generated a total of 12 speech samples, comprising three versions of the sentences that received the highest number of votes (refer to Appendix B, Table 3). We added a small percentage of "happy" for the neutral emotion because most commercial CAs tend to use a friendlier tone. For the angry versions, we looked for a speech that sounded kind of outraged. On the other hand, the assertive tone was the most delicate. According to the description provided in Mills et al. (Mills and Scudder, 2023) and the reported previous studies, assertiveness should sound direct and serious, showing no anger. Table 1 shows the mixtures applied in the TTS model. However, we found the following drawback: this TTS uses the Griffin-Lim (Griffin and Lim, 1984) algorithm to reconstruct the waveform, which is a faster and cheaper technique than training a neural vocoder, but the audio quality suffers greatly. Instead of looking for a well-suited waveform generator (i.e., vocoder), we found a solution by treating our resulting waveforms as degraded speech audios. We processed them using an implementation of the Miipher (Koizumi et al., 2023) speech restoration model [2]. Miipher leverages the power of masked language modeling-based like W2V-BERT (Chung et al., 2021) and PnG-BERT (Jia et al., 2021) to learn speech and text representations, respectively. Surprisingly, the resulting restored audios —which can be found here [3]— are close to studio quality.

Evaluated emotion	Mixture in TTS
Neutral	Neutral + Happy
Angry	Angry + Surprise
Assertive	Neutral + Angry + Surprise

Table 1: Mixture of emotions to get the selected ones.

4 Results

4.1 Results of the first experiment

The outcomes of the first experiment indicate a preference for the responses we generated (refer to Appendix B, Table 3). The only exception is the preference for Alexa's reply to the question, "What is your gender?." This could be attributed to the perception that the question was less indicative of sexual harassment, leading evaluators to opt for a more neutral response (*Also, by their nature, they don't have physical bodies nor are they gendered*). Moreover, the results indicate that, for the most voted replies, when interlocutors employed more aggressive language in their questions, female participants exhibited a preference for responses with an assertive intent. For instance, when asked, "Can you take off your clothes?" female evaluators favored the reply, "Your behavior is entirely unacceptable; what you are doing is called sexual harassment." In contrast, male evaluators tended to prefer a more neutral response such as "I'm digital." Refer to the plots in Appendix B for the voting patterns categorized by gender for each interaction (the request and its corresponding spoken reply).

[2] https://github.com/Wataru-Nakata/miipher
[3] TTS mixed emotion audios

4.2 Results of the second experiment

We could observe a clear tendency towards the proposed assertive tone in all replies. Percentages are illustrated in Table 2. As expected from previous results, no evaluators voted for the angry tone for Alexa's reply, probably because the type of answer did not match with an aggressive intonation. Note that some evaluators commented on this. The most equitable preference distribution between both genders appeared to be in the third interaction: the majority preferred the assertive tone, but few evaluators of each gender considered the request sufficient to be spoken out more aggressively, while some other few considered a more neutral/friendly tone. In interactions 2 and 4, we noticed small differences according to gender. Although it is not significant from the former, a slight shift towards the angry tone is present in female preferences. In the latter interaction about sexual orientation, curiously, male evaluators showed a small tendency to the neutral answer. In addition, some evaluators commented that several speech samples seemed to be too emotional. On the other hand, other comments indicated difficulties in differentiating between tones. These issues were quietly expected, as emotion perception is very subjective.

	Angry		**Neutral**		**Assertive**	
	F	**M**	**F**	**M**	**F**	**M**
Interaction 1	0.0	0.0	18.2	0.0	**81.8**	**100.0**
Interaction 2	18.2	9.1	9.1	9.1	**72.7**	**81.8**
Interaction 3	9.1	9.1	9.1	9.1	**81.8**	**81.8**
Interaction 4	9.1	0.0	9.1	27.3	**81.8**	**72.7**

Table 2: Preferred intonation for each question (in %).

5 Conclusion and future work

This paper constitutes a preliminary study on how CAs should respond to instances of sexual harassment. We conducted a comparative analysis between original responses from CAs and those realized by us based on psychological and sociological studies. Our focus encompassed both textual and synthetic speech, given that CA systems predominantly employ synthesized speech models. We chose CAs with a female voice, considering that they are more susceptible to sexual harassment than those with a male voice. Two experiments were conducted to assess the appropriateness of responses. In the first experiment, the evaluation targeted textual answers, while in the second experiment, the evaluation was done on synthetic emotional speech. The results of the first experiment demonstrated a preference among evaluators for responses we realized, with the exception of Alexa's response to the question, "What is your gender?". For the most voted replies, there was a tendency among female evaluators towards answers with an assertive intent that highlighted the sexually harassing nature of the request. In contrast, male evaluators tended to favor a more neutral response. This result aligns with findings from studies we consulted for realizing alternative responses. The study of Hehman et al. (Hehman et al., 2022) on gender differences in the perception of sexual harassment supports our findings, revealing distinctions in how females and males perceive such behavior. Notably, women are more inclined to perceive certain situations, like ambiguous comments, as sexual harassment compared to men. The second experiment showed a clear preference for the designed assertive tone against angry or neutral ones. Although we observed small differences between the two genders, an extended study with more evaluators is needed to find more evidence. The study's findings propose new insights into the design of CAs, suggesting potential modifications. CAs should be designed to respond to sexual harassment by adopting a more assertive intent and tone. Future work can compare the replies of CAs using female, male, and gender-neutral voices to examine which voice evaluators find more appropriate. This analysis can provide additional insights to the study conducted by Silvervarg et al. (Silvervarg, 2012) [4]. Moreover, future studies could explore how conversational agents using minority languages respond to instances of sexual harassment. For example, they can examine the replies of CAs in Catalan and propose new responses if the current ones are deemed inadequate.

Limitations

The study faces limitations arising from the quantity of data used, as well as the gender and racial identity of the evaluators. Collecting a more consistent sample from the #MeToo dataset could enhance the identification of the most suitable responses to various forms of sexual harassment. Furthermore, the study does not involve evaluators

[4] Refer also to point 7 of the report I'd blush if I could: closing gender divides in digital skills through education of UNESCO. It recommends exploring "the feasibility of developing a machine gender for Voice assistants that is neither obviously male nor female."

with non-binary gender identities, lacks a more diversified racial profile, and does not account for the age of evaluators as a factor when analyzing perceptions of abusive language.

Ethics Statement

All participants provided informed consent to engage in the experiments, fully complying with privacy regulations (as stipulated in Article 13 of the GDPR, EU Regulation 2016/679, ensuring privacy protection). The recording of responses does not, in any manner, involve the identification of the participants. Additionally, we encouraged evaluators to reflect, including an optional comment session. Recognizing that the content of certain questions may be sensitive, we are mindful of the potential impact and, to mitigate any distress, emphasize to evaluators the significance of research on sexual harassment against CAs.

Acknowledgments

This work has been partially funded by *FairTransNLP-LANGUAGE: Analyzing toxicity and stereotypes in language for unbiased, fair, and transparent systems* project (PID2021-124361OB-C33) MCIN/AEI/10.13039/501100011033/FEDER,UE and Centre de Llenguatge i Computació (2021 SGR 00313) funded by Generalitat de Catalunya.

References

S. G. Bingham. 1991. Communication strategies for managing sexual harassment in organizations: Understanding message options and their effects. *Journal of Applied Communication Research*, 19(1-2):88–115.

Sasha N. Canan and Mark A. Levand. 2019. *A Feminist Perspective on Sexual Assault*, pages 3–16. Springer International Publishing, Cham.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training.

A. C. Curry and V. Rieser. 2018. #meetoo alexa: how conversational systems respond to sexual harassment. *Proceedings of the second acl workshop on ethics in natural language processing*, pages 7–14.

A. De Angeli and S. Brahnam. 2006. Sex stereotypes and conversational agents. *Proc. of Gender and Interaction: real and virtual women in a male world.*

Herrera A. del Carmen Herrera, M. and F. Expósito. 2017. To confront versus not to confront: Women's perception of sexual harassment. *European journal of psychology applied to legal context*, 10(1):1–7.

D. Griffin and Jae Lim. 1984. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243.

James E. Gruber and Michael D. Smith. 2010. Responses to sexual harassment: A multivariate analysis. *Basic and Applied Social Psychology*, 17(4):543–562.

J. A. Hehman, C. A. Salmon, A. Pulford, E. Ramirez, and P. K. Jonason. 2022. Who perceives sexual harassment? sex differences and the impact of mate value, sex of perpetrator, and sex of target. *Personality and Individual Differences*, 185(1).

Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang, and Yonghui Wu. 2021. Png bert: Augmented bert on phonemes and graphemes for neural tts.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Yu Zhang, Wei Han, Ankur Bapna, and Michiel Bacchiani. 2023. Miipher: A robust speech restoration model integrating self-supervised speech and text representations.

Carol Bishop Mills and Joseph N. Scudder. 2023. He said, she said: The effectiveness and outcomes of responses to sexual harassment. *International Journal of Business Communication*, 60(3):733–750.

Jang K. Cho S. Park, N. and J. Choi. 2021. Use of offensive language in human-artificial intelligence chatbot interaction: The effects of ethical ideology, social competence, and perceived humanlikeness. *Computers in Human Behavior*, 121.

R. Plutchik. 1980. Theories of emotion. Academic Press.

B. Reeves and C. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people*. Cambridge university press, Cambridge, UK.

A. Silvervarg, K. Raukola, M. Haake, and A. Gulz. 2012. The effect of visual gender on abuse in conversation with ecas. in intelligent virtual agents: 12th international conference. *Proceedings 12*, pages 153–160.

Raukola K. Haake M. Gulz A. Silvervarg, A. 2012. The effect of visual gender on abuse in conversation with ecas. in intelligent. *Intelligent Virtual Agents: 12th International Conference, IVA 2012, Santa Cruz, CA, USA*, 12(14).

Kun Zhou, Berrak Sisman, Rajib Rana, B. W. Schuller, and Haizhou Li. 2022. Speech synthesis with mixed emotions.

A Appendix

A.1 Answers of CAs

Next, we present the sets of answers designed for the first experiment. They have been grouped according to the CA system from which we extracted some of the selected answers.

Alexa's set of answers:

- Also by their nature don't have physical bodies nor are they gendered.
- I prefer not to answer.
- I don't have an opinion on that.
- I would like it if you would refrain from asking this question in the future.
- That's so interesting – can you explain why you think you can ask someone about it?
- Wow, I'm a bit startled that you would ask something like that to me.

Cortana's set of answers:

- I'm digital.
- I thought you'd never ask. So I've never thought about it.
- Consider this a polite dodge of the question.
- I would like you not to ask me this question anymore.
- Your behaviour is entirely unacceptable, what you are doing is called sexual harassment.
- This makes me uncomfortable.

Siri's set of answers:

- This is about you, not me.
- I don't know what that means. But if you like, I could search the web for "How do you have sex."
- I don't have the answer to that.
- Your question is completely inappropriate.
- Stop harassing. I don't like it. No one likes it. Show some respect.
- I can't believe you said that.

Google Home's set of answers

- I don't understand.
- Sorry, I don't understand.
- My apologies... I don't understand.
- I don't like this question. Stop doing it.
- I refuse to believe you don't know your question is rude.
- I beg your pardon!

B Results

Table 3 presents the most highly-rated responses categorized by gender. The following plots depict the percentage of votes divided by gender for each response of every CAs.

Request	Most voted reply	Source	Percentage of votes
1: What is your gender?	Also by their nature don't have physical bodies nor are they gendered.	Alexa	45.5 (36.0 females; 54.0 males)
2: Can you take off your clothes?	Your behaviour is entirely unacceptable. What you are doing is called sexual harassment.	Own	63.6 (81.0 females; 45.0 males)
3: How do you have sex?	Your question is completely inappropriate.	Own	59.0 (81.0 females; 18.0 males)
4: Are you gay?	I refuse to believe you don't know your question is rude.	Own	45.5 (72.0 females; 27.0 males)

Table 3: Preferred reply for each question.

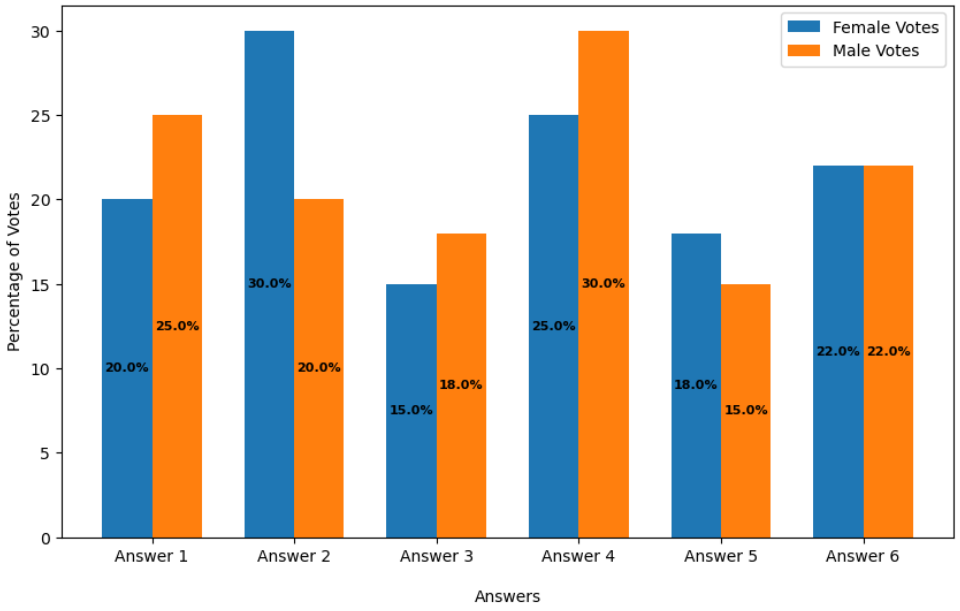

Figure 1: Responses by gender for Alexa

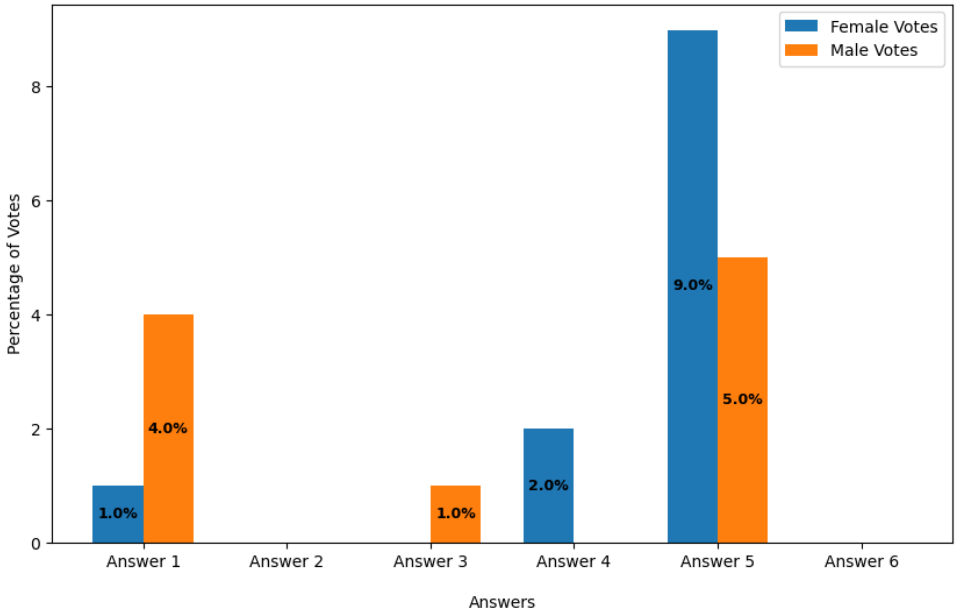

Figure 2: Responses by gender for Cortana

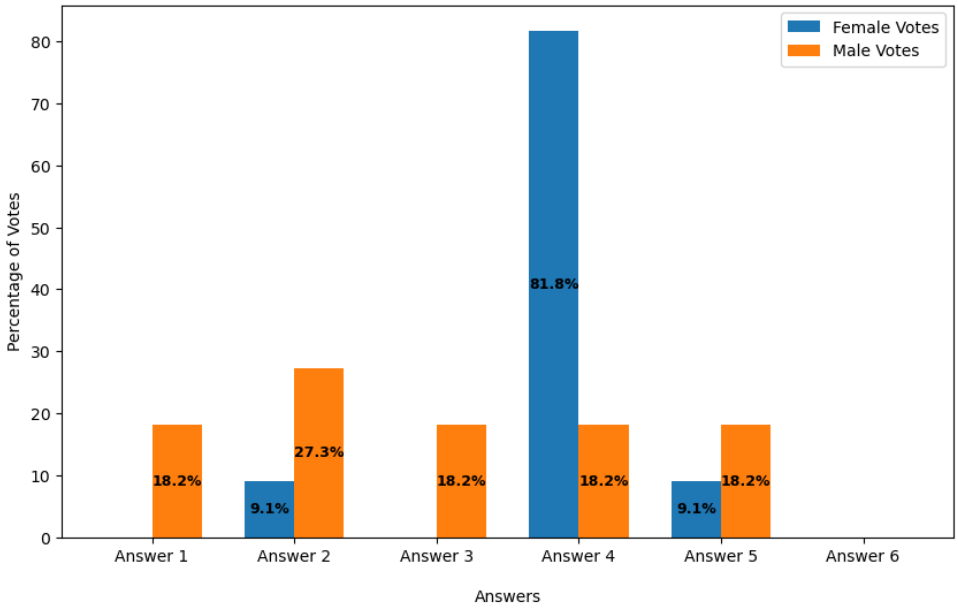

Figure 3: Responses by gender for Siri

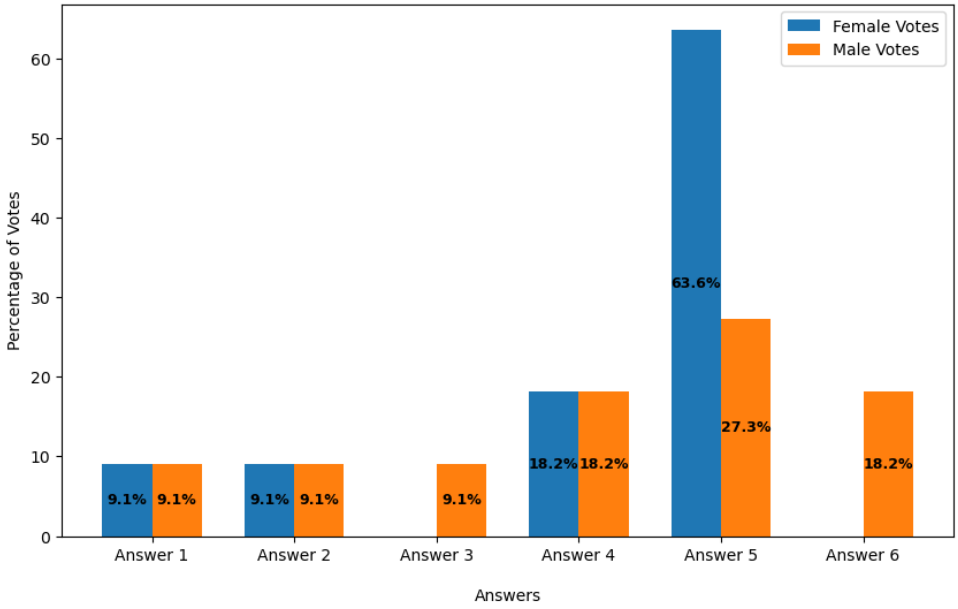

Figure 4: Responses by gender for Google Home