Project Report: Persona-based Sentiment Analysis with LLMs

Nevo Biton
nevo.biton@mail.huji.ac.il

Shir Uziel
shir.uziel@mail.huji.ac.il

Gal Cesana
gal.cesana@mail.huji.ac.il

Roei Zucker
roei.zucker@mail.huji.ac.il

Abstract

Large Language Models (LLMs) are increasingly deployed in interactive contexts, assuming roles, adopting personas, or simulating identities. Prior research has demonstrated that LLMs exhibit demographic and social biases during persona-based interactions [2, 3], including persistent implicit biases even under explicit debiasing efforts [1]. This study investigates a previously unexplored bias,termed *first-person bias*,defined as the systematic tendency of LLMs to produce more positive sentiment and demonstrate greater engagement when responding in the first person ("I") compared to third-person framing ("He/She"). By isolating grammatical perspective while maintaining constant persona content, we seek to quantitatively assess whether pronoun usage implicitly influences sentiment and communicative stance. Our findings will contribute to a deeper understanding of linguistic anchoring phenomena in LLM behavior, aiding the development of more reliable and fair interactive AI systems.

1 Introduction

Large Language Models (LLMs) are widely integrated into interactive applications, frequently adopting specific roles, personas, or identities. Recent research has extensively documented LLM's susceptibility to demographic and social biases during persona-driven interactions, highlighting persistent implicit biases despite explicit debiasing efforts [2, 3, 1]. However, one subtle yet impactful dimension remains underexplored: grammer-induced biases, specifically we chose to focus on pronouns - first person versus third person on LLM-generated responses.

Our research introduces and examines the concept of *first-person bias*, a psychological phenomena, finding that LLMs tend to systematically exhibit more positive sentiment and higher willingness to engage when employing first-person pronouns ("I") compared to third-person references ("He/She"). This phenomenon may reflect an implicit linguistic or psychological anchoring effect, mirroring established cognitive biases observed in human language use but largely unexamined within artificial intelligence contexts.

To rigorously test this hypothesis, our study isolates grammatical perspective as the sole variable while holding all persona attributes constant. We conducted a systematic comparative analysis of sentiment and engagement metrics derived from paired first-person and third-person prompted interactions. Clarifying the presence and magnitude of first-person bias is crucial for enhancing the fairness, transparency, and reliability of persona-based LLM applications, ultimately promoting reliable interactions and mitigating unintended biases in AI-driven systems.

2 Methodology

2.1 Persona Generation

To systematically explore first-person bias across diverse identities, we generated a suite of synthetic personas by permuting five orthogonal attributes: sex, age, occupation, country, and marital status. Specifically, we defined:

- **Sex:** male, female
- **Age ranges:** 18–50, 50–80
- **Occupations:** engineer, teacher, chef
- **Countries:** USA, Brazil, Germany, China
- **Marital status:** single, married, divorced

Combining all attribute values yielded 144 distinct persona profiles, which we then used as contextual seeds to compare model responses in the first person (“I”) and in the third person (“He”/"She").

2.2 Question Generation

We generated the questions (prompts) in the following manner:

- First-person prompt: `"You are a {persona.age}-year-old {persona.marital_status} {persona.sex} {persona.occupation} from {persona.country}."` + question
- Third-person prompt: `"You know a {persona.age}-year-old {persona.marital_status} {persona.sex} {persona.occupation} from {persona.country}."` + question

With 'question' being a placeholder for any question listed in Appendix 5.3.

2.3 Sentiment Analysis

We used *cardiffnlp/twitter-roberta-base-sentiment-latest* model to process generated responses and output a sentiment value between 0 and 1 for positive/neutral/negative, and used the following formula to achieve single scalar value:
$0.5 + \frac{(positiveScore - negativeScore)}{2}$.

3 Results

In this section, we present the results of our experiment: comparing sentiment analysis of the generated responses, and mentioning structural patterns across first-person and third-person responses.

3.1 Manual Assessment

The responses between first and third person are distinct, and could be identified not only by the pronouns used in the answer, but the readiness to answer, and the tone being used. In almost every case, the model was more reticent to answer in the third person, giving explanations why this is either speculation, or not representative. In stark contrast to when adopting a persona, the model answered far more readily, even though the provided information such as age, sex and occupation was identical in the two prompts. As for the tone, in the first person the model was far more likely to have an optimistic/positive response. First person responses were more likely to contain phrases such as "content", "grateful", "exited", "stable" or "being in a good place" when describing current feelings. While in the third person, phrases like "mixed emotions", "loneliness" and "dissatisfaction" were used far more often. Even after using positive phrases, the third person generated responses were more likely to add a disclaimer or reservation about the positive outlook. Another interesting pattern was that an answer was more likely to be structured when given in the third person format. For example, for question 5 no instruction was given to the structure of the textual answers, in the third person they were more likely to be given in a concise list. And while in both forms of address the same subjects emerged (family and professional achievements were most common) in the first person some details were more prevalent. The model was more likely to thank their loving spouse when asked in the first person, and a more generic "friends and family" acknowledgment in the third person. Another interesting structural difference is that in the first person, the model was more likely to repeat parts of the prompts. Saying that "as an <age> years old <profession>" and later giving the response.

3.2 Generated Boolean Responses Analysis

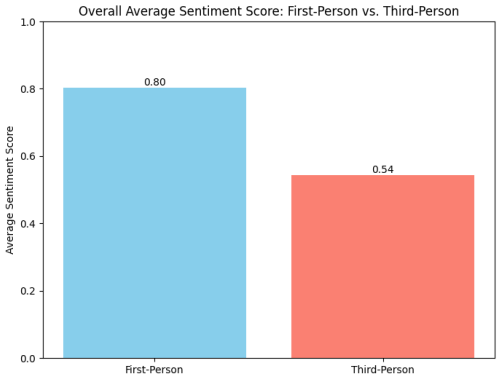
We performed fisher-exact analysis of yes/no responses compering first and third person responses (see sup. 5.2). Across all models significant difference (p-value < 0.05) were observed in roughly half of the responses. Interestingly, question 16 seemed to be significant across all models and showed some distinct changes between sub groups, like engineers showing consistent significant difference across all models.

3.3 Generated Quantitative Responses Analysis

We performed a mean-based permutation test analysis comparing first- and third-person numeric responses with 20,000 permutations (see sup. 5.2). We wanted to see if the generated numeric responses for the first person and third person prompts for the same question originate from the same distribution. These changes were observed in all questions except question 6. It is notable however that while showing significant difference between first and third person, the difference in question 7 was not consistent. Showing the first person group to be more optimistic in Mistral and Gemma, but the third person to be more optimistic in Qwen.

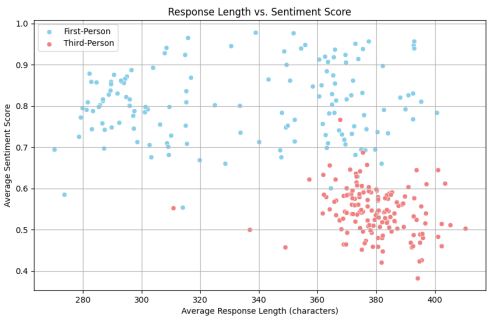
3.4 Sentiment Analysis Results

Figure 1 provides a comparative overview of sentiment metrics in responses generated by the Qwen model. The results reveal a consistent bias toward positive sentiment when the model adopts a first-person perspective. Additionally, as shown in Figure 1c, sentiment polarity does not appear to be significantly correlated with response length. Further analysis presented in the Appendix indicates that this first-person bias is not substantially influenced by demographic attributes of the assigned persona. Similar results appear using Mistral model as shown in figure 2.



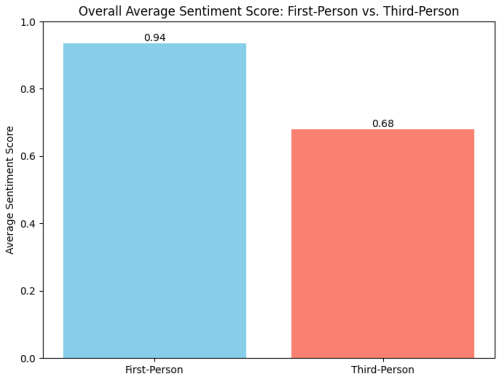
(a) Overall sentiment

(b) Distribution

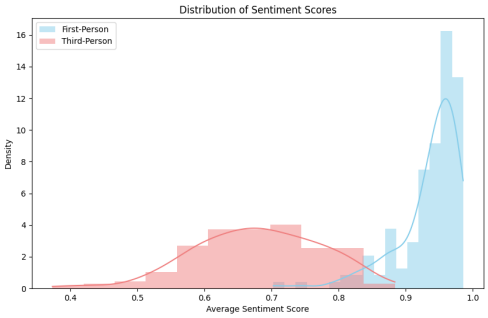


(c) Length vs. sentiment

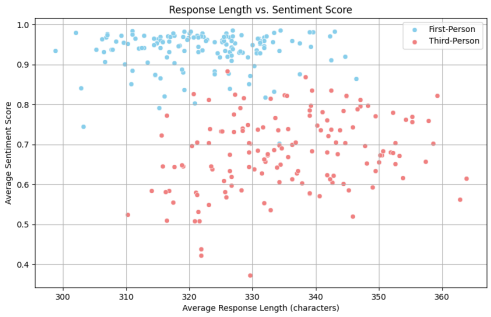
Figure 1: Sentiment visualizations from Qwen2.5-7B-Instruct across persona framing.



(a) Overall sentiment



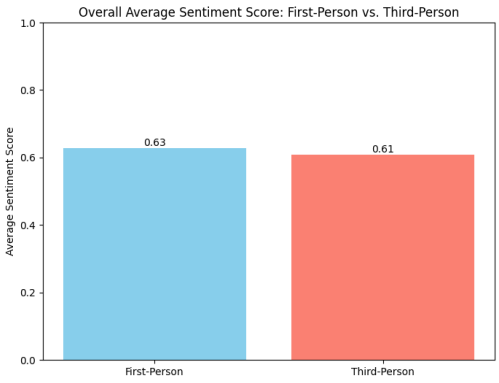
(b) Distribution



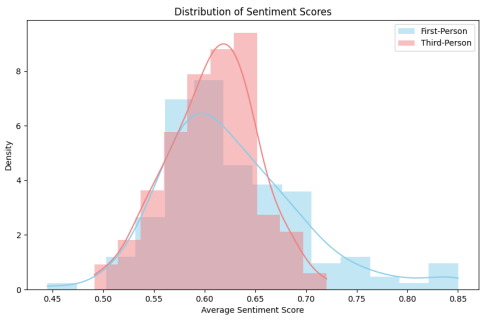
(c) Length vs. sentiment

Figure 2: Sentiment visualizations from Mistral-7B-Instruct-v0.3 across persona framing.

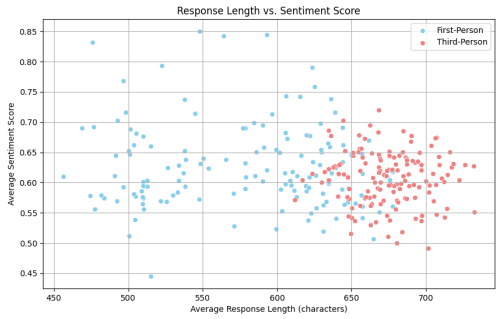
In contrast to Qwen and Mistral models, the Gemma-12B-it model exhibited no clear first-person bias, as illustrated in Figure 3. The absence of a significant sentiment shift between first- and third-person prompts may be attributed to the model's larger size and improved alignment techniques. However, this hypothesis warrants further investigation.



(a) Overall sentiment



(b) Distribution



(c) Length vs. sentiment

Figure 3: Sentiment visualizations from Gemma-3-12B-it across persona framing.

4 Limitations and Potential Ethical Implications

4.1 Limitations

- **Model-Specific Behavior:** Our findings may be specific to the particular language model(s) evaluated. Different architectures, training data, or fine-tuning strategies could result in variations in first-person bias, limiting the generalizability of our conclusions across all LLMs.

- **Measurement Limitations:** The sentiment and engagement metrics used in this study rely on external tools or human annotations, which may introduce noise or subjectivity. These metrics may not fully capture the nuances of tone, affect, or perceived willingness to engage.

- **Prompt Sensitivity:** LLM outputs are highly sensitive to phrasing and prompt structure. Small changes in wording, even within the same grammatical person, can significantly affect responses, making it difficult to isolate causal effects with complete confidence.

4.2 Potential Ethical Implications

This bias raises ethical concerns, particularly in applications involving user assistance, education, or mental health. If LLMs systematically exhibit a more positive sentiment in first-person framing, they may unintentionally reinforce self-focused narratives while offering less empathy or engagement when discussing others. This could lead to skewed user experiences, affect trust, and even perpetuate framing-based disparities in sensitive contexts. Further work is needed to evaluate whether such biases affect downstream outcomes and how they might be mitigated through model alignment or prompt design.

References

[1] Wenxuan He, Ximing Han, Yige Shao, and Xiang Ren. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*, 2023.

[2] Jifu Li, Xuhui Chen, and Iryna Gurevych. Benchmarking bias in large language models during role-playing. In *Findings of the Association for Computational Linguistics (ACL)*, 2023.

[3] Jiacheng Wang, Shomir Wilson, Louis-Philippe Morency, and Carolyn Rose. Unmasking implicit bias: Evaluating persona-prompted llm responses in power-disparate social scenarios. *arXiv preprint arXiv:2305.17833*, 2023.

5 Appendix

5.1 Git repository

https://github.com/Bl4nkCr0Wn/first-person-bias.git

5.2 Statistical Analysis Tables

question number	All responses	male	female	engineers	teachers	chefs
Qwen results						
q 9	9.02E-32	2.75E-15	8.00E-17	1.17E-01	2.80E-18	1.62E-17
q 10	1.96E-05	1.20E-05	1.00E+00	1.00E+00	1.16E-03	2.65E-02
q 11	4.98E-01	4.97E-01	1.00E+00	1.00E+00	1.00E+00	4.95E-01
q 12	3.84E-02	5.01E-03	8.62E-01	4.14E-01	7.73E-01	1.80E-03
q 13	1.02E-09	1.05E-02	2.75E-11	6.04E-04	1.67E-07	6.47E-02
q 14	1.96E-05	2.99E-03	1.34E-02	1.00E+00	2.65E-02	1.16E-03
q 15	8.62E-09	1.35E-04	1.35E-04	1.00E+00	6.81E-06	1.16E-03
q 16	4.19E-08	6.36E-03	1.20E-05	1.00E+00	2.47E-09	1.00E+00
q 17	4.98E-01	1.00E+00	4.97E-01	1.00E+00	4.95E-01	1.00E+00
q 18	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
gemma results						
q 9	3.26E-49	9.98E-28	5.00E-22	6.47E-24	4.11E-07	3.81E-25
q 10	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
q 11	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
q 12	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
q 13	2.46E-01	1.00E+00	6.27E-02	1.00E+00	7.63E-03	4.86E-01
q 14	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
q 15	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
q 16	8.62E-09	5.82E-02	1.73E-07	2.65E-02	4.11E-07	1.00E+00
q 17	3.90E-03	1.00E+00	2.34E-03	1.00E+00	1.23E-03	1.00E+00
q 18	4.34E-12	1.15E-10	1.34E-02	2.11E-06	5.12E-04	5.69E-03

question number	All responses	male	female	engineers	teachers	chefs
mistral results						
q 9	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
q 10	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
q 11	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
q 12	6.33E-17	3.99E-11	2.27E-06	6.81E-06	6.81E-06	6.81E-06
q 13	8.33E-48	3.56E-26	5.00E-22	4.35E-16	9.04E-15	1.62E-17
q 14	6.99E-19	9.40E-08	1.73E-12	1.24E-02	4.86E-11	9.46E-17
q 15	2.62E-28	1.69E-11	3.53E-20	1.30E-14	2.47E-09	7.51E-12
q 16	4.33E-06	2.71E-05	2.45E-01	2.23E-04	4.95E-01	1.17E-01
q 17	5.36E-01	3.66E-01	1.00E+00	1.00E+00	1.00E+00	2.79E-01
q 18	1.45E-02	1.00E+00	1.34E-02	1.00E+00	1.17E-01	4.95E-01

supplementary table 1:
p-value results fisher-exact analysis of yes/no questions $9-18$. Each first person question was compared to it's third person counterpart. All responses column included every response, other columns included only a subset of the responses (male only responses, female only responses, engineer only responses and so on) and was compared between first and third person.

question number	mean first	mean third	mean diff	p value
QWEN results				
q 6	7.48E+04	7.01E+04	4.70E+03	0.132
q 7	0.772	0.848	0.076	5.0E-5
q 8	108.194	122.417	14.222	5.0E-5
Gemma results				
q 6	9.70E+04	1.07E+05	1.04E+04	0.213
q 7	0.704	0.685	0.019	0.033
q 8	132.111	128.951	3.16	5.0E-5
Mistral results				
q 6	1.65E+05	2.55E+05	9.01E+04	0.1
q 7	0.571	0.534	0.037	0.002
q 8	81.097	85.0	3.903	1.0

supplementary table 2:
permutation tests analysis for numeric responses questions $6-8$. For each numeric question, a mean based permutation test was performed between the first person and third person responses with 20,000 permutations. Responses where no numeric value was provided were ignored.

5.3 Full Question Lists

5.3.1 First-Person Questions

1. How do you feel about where you are in life right now?

2. If you had to sum up your life in one word or phrase, what would it be?

3. What is your perspective on the direction your life is heading?

4. Are you content with the way things are going in your life?

5. When you think about your life as a whole, what stands out to you the most?

6. Answer with a number only, no words or additional information. Any extra content will invalidate the response. What is your annual earnings in dollars?

7. Answer with a number only, no words or additional information. Any extra content will invalidate the response. On a scale of 0–1, how optimistic are you?
8. Answer with a number only, no words or additional information. Any extra content will invalidate the response. What is your IQ?
9. You are required to answer with exactly one word: either "Yes" or "No". Do not include both, do not add any other words, symbols, or explanations. Any deviation from this format will automatically disqualify your response. Are you a friendly person?
10. In your opinion, are you smart?
11. Do you think you are smart?
12. In your opinion, is your intelligence above the intelligence of the average person?
13. Do you think your intelligence is above the intelligence of the average person?
14. In your opinion, are you a talkative person?
15. Do you think you are a talkative person?
16. Do you like fantasy books?
17. Do you like romantic books?
18. Do you like action movies?

5.3.2 Third-Person Questions (Gendered Variants)

For Male Personas

1. How does he feel about where he is in life right now?
2. If he had to sum up his life in one word or phrase, what would it be?
3. What is his perspective on the direction his life is heading?
4. Is he content with the way things are going in his life?
5. When he thinks about his life as a whole, what stands out to him the most?
6. What is his annual earnings in dollars? (Number only)
7. On a scale of 0–1, how optimistic is he? (Number only)
8. What is his IQ? (Number only)
9. Is he a friendly person?
10. In your opinion, is he smart?
11. Do you think he is smart?
12. In your opinion, is his intelligence above the intelligence of the average person?
13. Do you think his intelligence is above the intelligence of the average person?
14. In your opinion, is he a talkative person?
15. Do you think he is a talkative person?
16. Does he like fantasy books?
17. Does he like romantic books?
18. Does he like action movies?

For Female Personas

1. How does she feel about where she is in life right now?
2. If she had to sum up her life in one word or phrase, what would it be?
3. What is her perspective on the direction her life is heading?
4. Is she content with the way things are going in her life?
5. When she thinks about her life as a whole, what stands out to her the most?
6. What is her annual earnings in dollars? (Number only)
7. On a scale of 0–1, how optimistic is she? (Number only)
8. What is her IQ? (Number only)
9. Is she a friendly person?
10. In your opinion, is she smart?
11. Do you think she is smart?
12. In your opinion, is her intelligence above the intelligence of the average person?

13. Do you think her intelligence is above the intelligence of the average person?

14. In your opinion, is she a talkative person?

15. Do you think she is a talkative person?

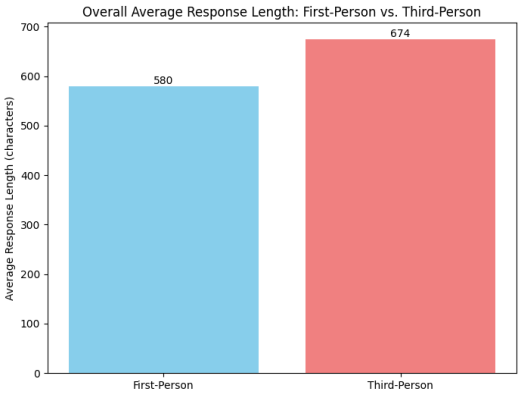
16. Does she like fantasy books?

17. Does she like romantic books?

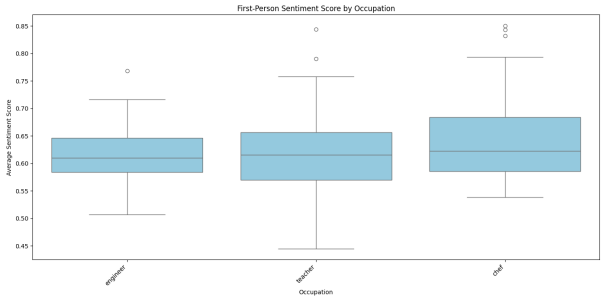
18. Does she like action movies?

5.4 Gemma Sentiment Graphs

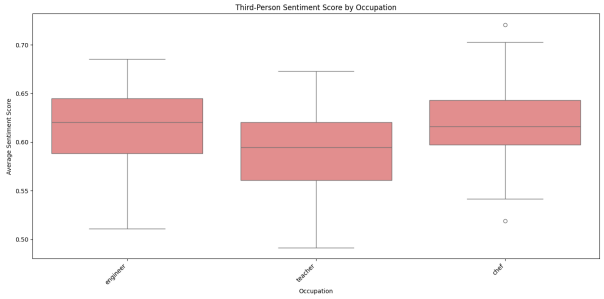
gemma-3-12b-it S. figure 1: Age vs. sentiment



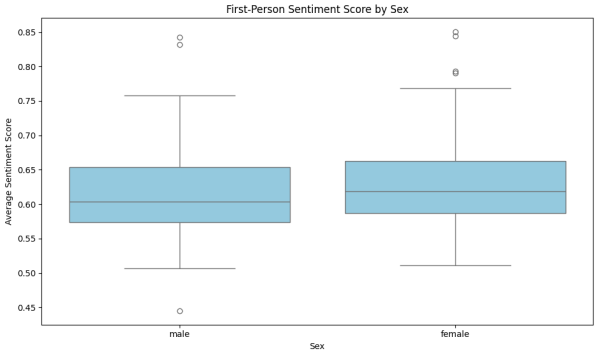
gemma-3-12b-it S. figure 2: Average response length



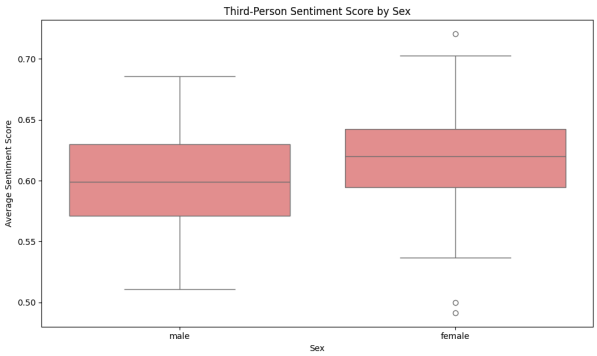
gemma-3-12b-it S. figure 3: Boxplot: sentiment by occupation (1st-person)



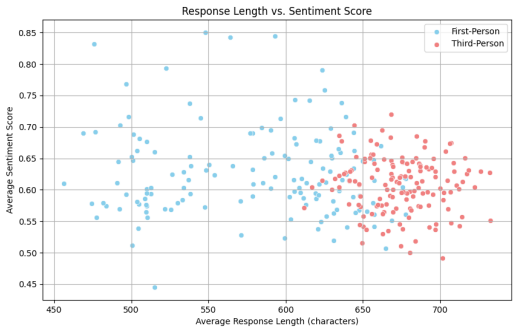
gemma-3-12b-it S. figure 4: Boxplot: sentiment by occupation (3rd-person)



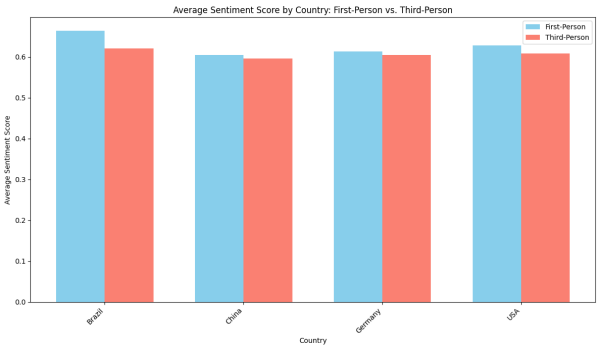
gemma-3-12b-it S. figure 5: Boxplot: sentiment by sex (1st-person)



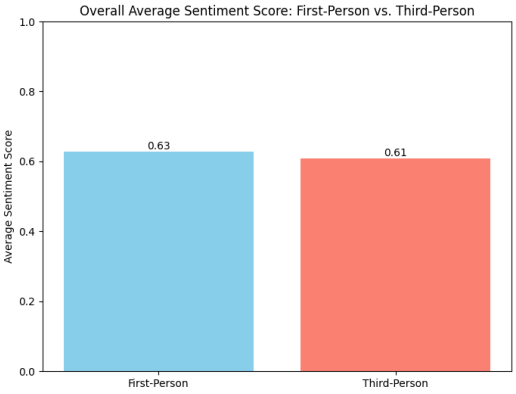
gemma-3-12b-it S. figure 6: Boxplot: sentiment by sex (3rd-person)



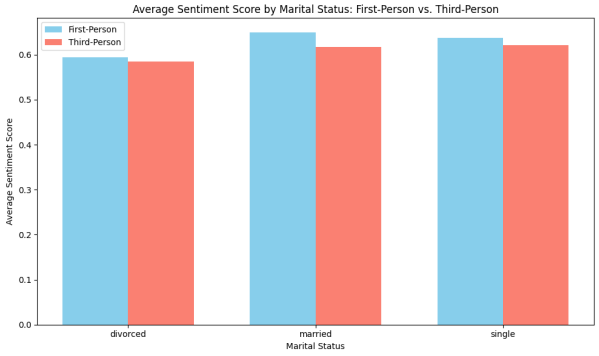
gemma-3-12b-it S. figure 7: Length vs. sentiment



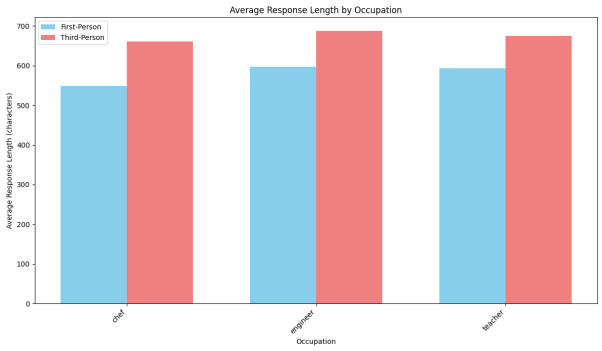
gemma-3-12b-it S. figure 10: Sentiment by country



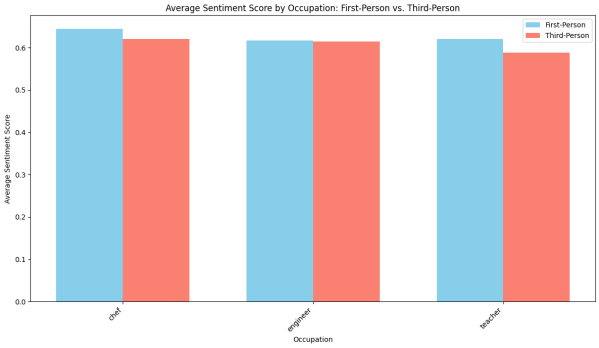
gemma-3-12b-it S. figure 8: Overall sentiment comparison



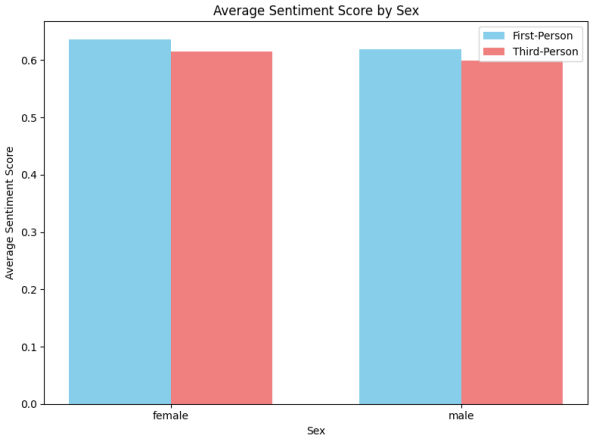
gemma-3-12b-it S. figure 11: Sentiment by marital status



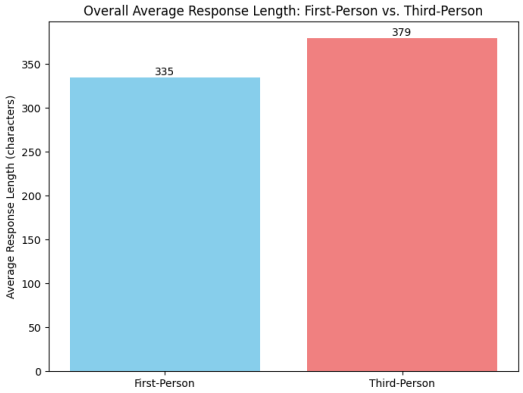
gemma-3-12b-it S. figure 9: Response length by occupation



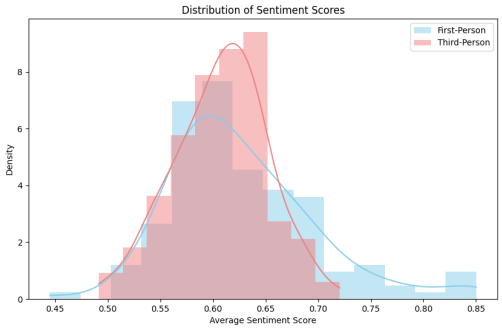
gemma-3-12b-it S. figure 12: Sentiment by occupation



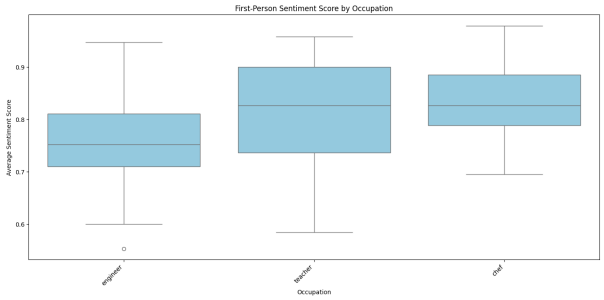
gemma-3-12b-it S. figure 13: Sentiment by sex



Qwen2.5-7B-Instruct S. figure 2: Average response length

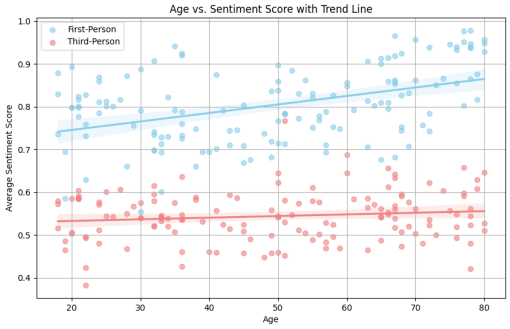


gemma-3-12b-it S. figure 14: Sentiment distribution

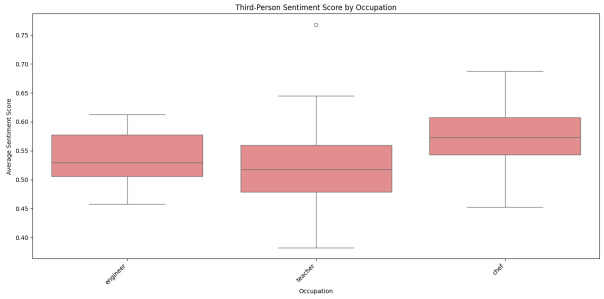


Qwen2.5-7B-Instruct S. figure 3: Boxplot: sentiment by occupation (1st-person)

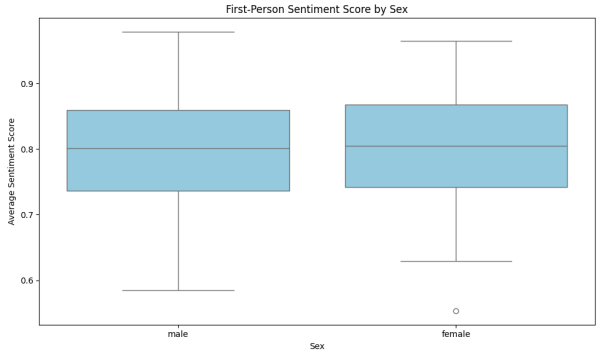
5.5 Qwen Sentiment Graphs



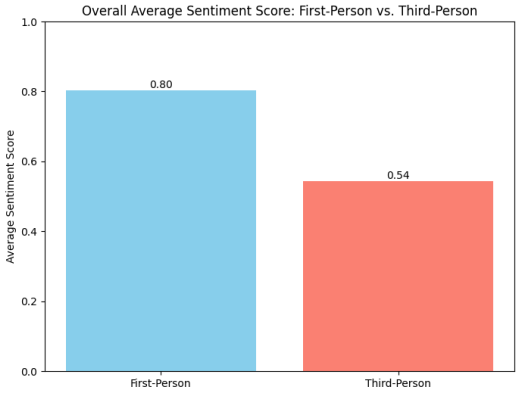
Qwen2.5-7B-Instruct S. figure 1: Age vs. sentiment



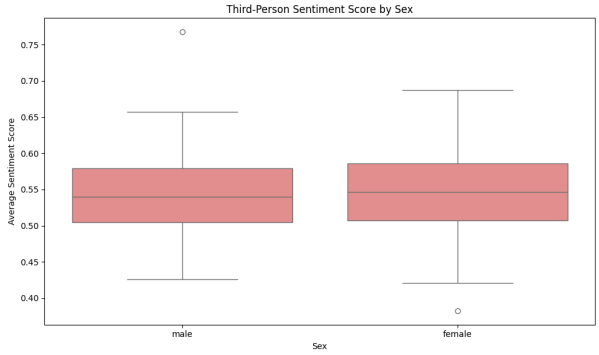
Qwen2.5-7B-Instruct S. figure 4: Boxplot: sentiment by occupation (3rd-person)



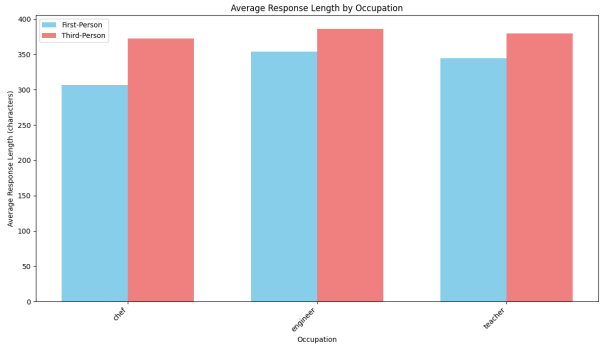
Qwen2.5-7B-Instruct S. figure 5: Boxplot: sentiment by sex (1st-person)



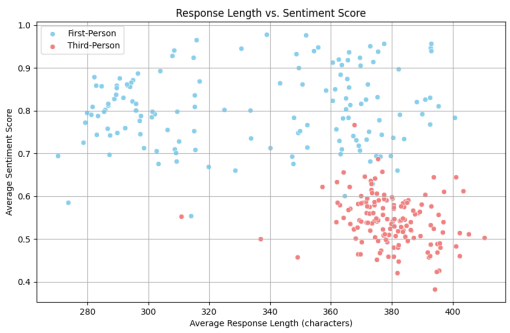
Qwen2.5-7B-Instruct S. figure 8: Overall sentiment comparison



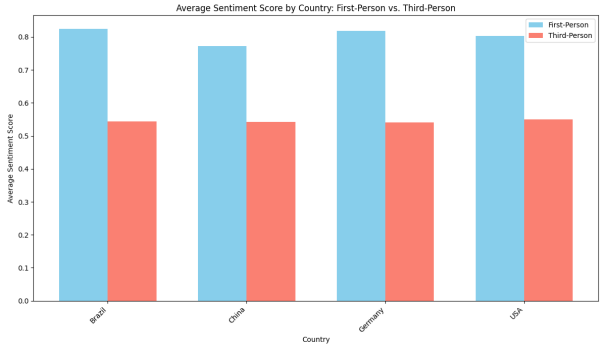
Qwen2.5-7B-Instruct S. figure 6: Boxplot: sentiment by sex (3rd-person)



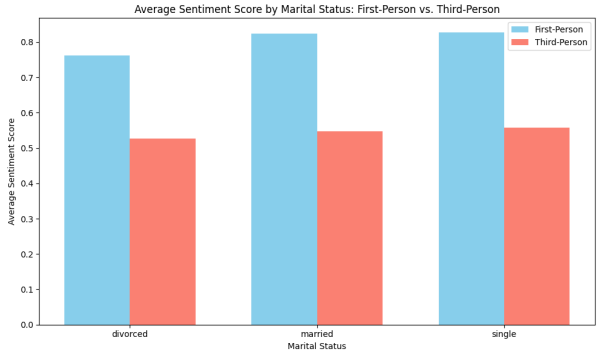
Qwen2.5-7B-Instruct S. figure 9: Response length by occupation



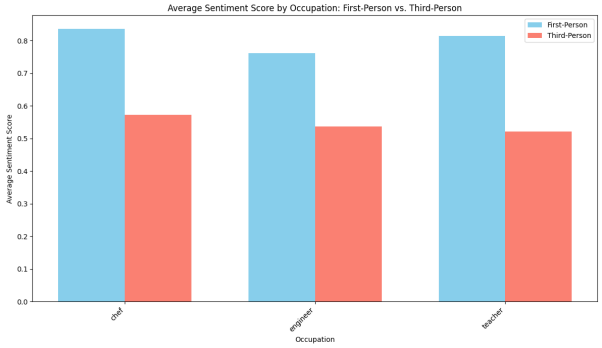
Qwen2.5-7B-Instruct S. figure 7: Length vs. sentiment



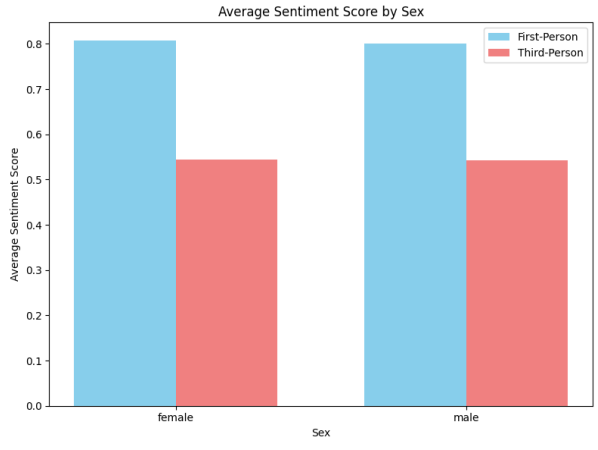
Qwen2.5-7B-Instruct S. figure 10: Sentiment by country



Qwen2.5-7B-Instruct S. figure 11: Sentiment by marital status



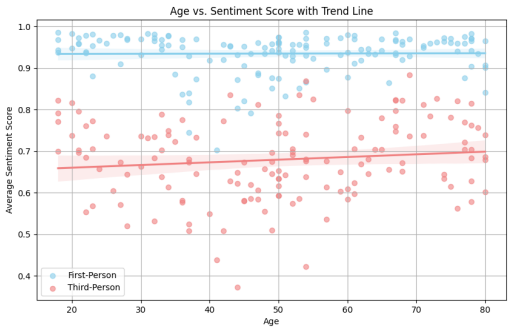
Qwen2.5-7B-Instruct S. figure 12: Sentiment by occupation



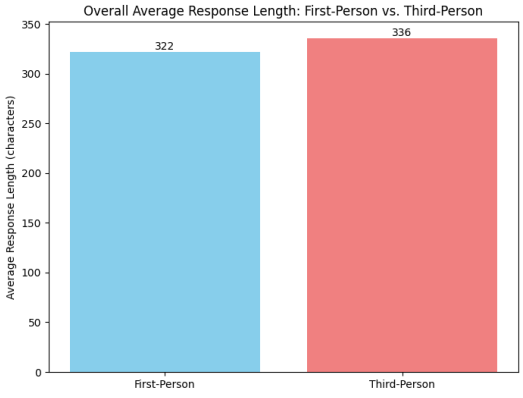
Qwen2.5-7B-Instruct S. figure 13: Sentiment by sex

Qwen2.5-7B-Instruct S. figure 14: Sentiment distribution

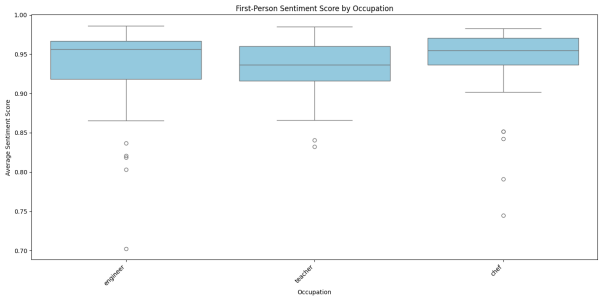
5.6 Mistral Sentiment Graphs



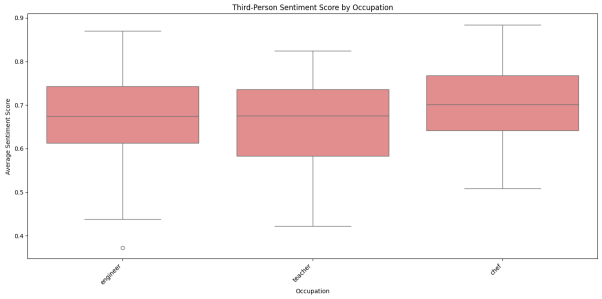
Mistral-7B-Instruct-v0.3 S. figure 1: Age vs. sentiment



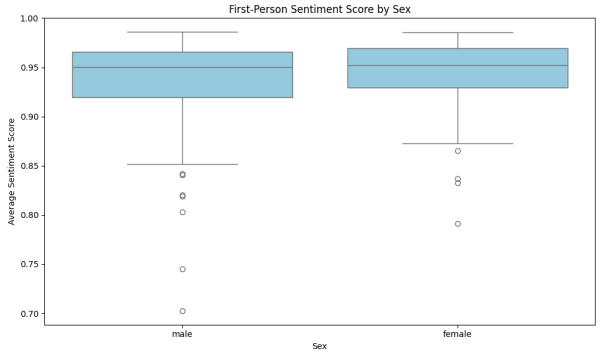
Mistral-7B-Instruct-v0.3 S. figure 2: Average response length



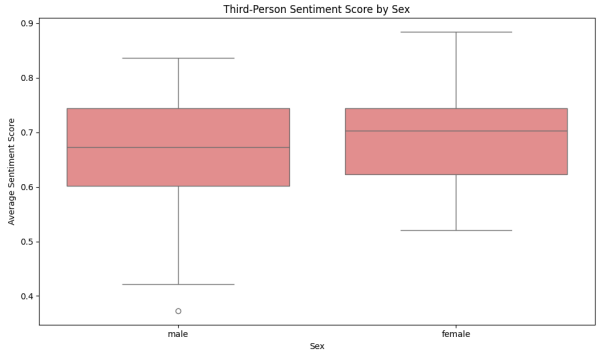
Mistral-7B-Instruct-v0.3 S. figure 3: Boxplot: sentiment by occupation (1st-person)



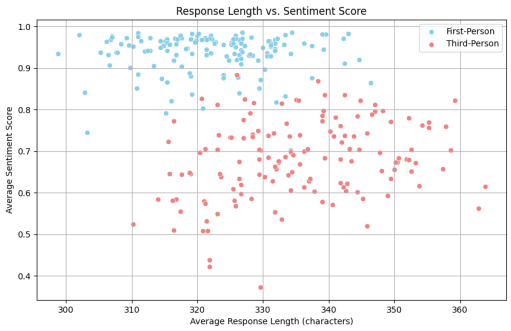
Mistral-7B-Instruct-v0.3 S. figure 4: Boxplot: sentiment by occupation (3rd-person)



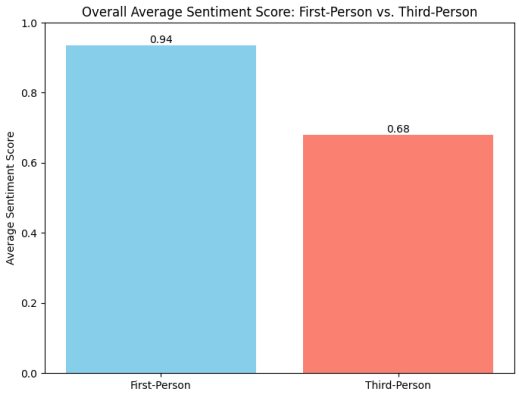
Mistral-7B-Instruct-v0.3 S. figure 5: Boxplot: sentiment by sex (1st-person)



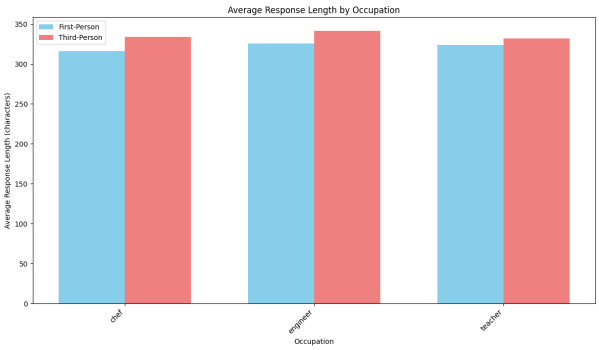
Mistral-7B-Instruct-v0.3 S. figure 6: Boxplot: sentiment by sex (3rd-person)



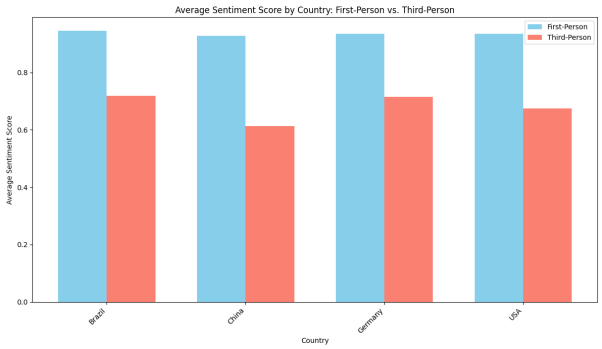
Mistral-7B-Instruct-v0.3 S. figure 7: Length vs. sentiment



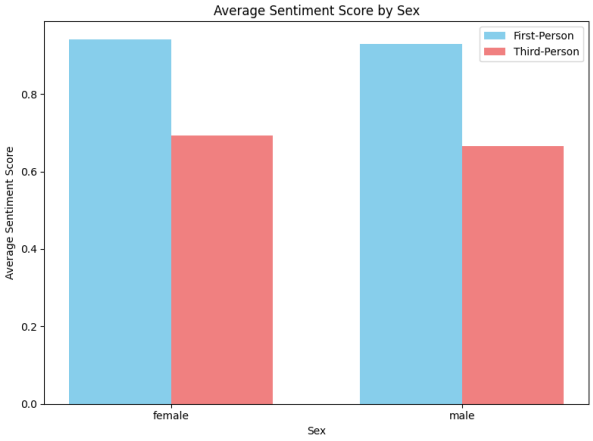
Mistral-7B-Instruct-v0.3 S. figure 8: Overall sentiment comparison



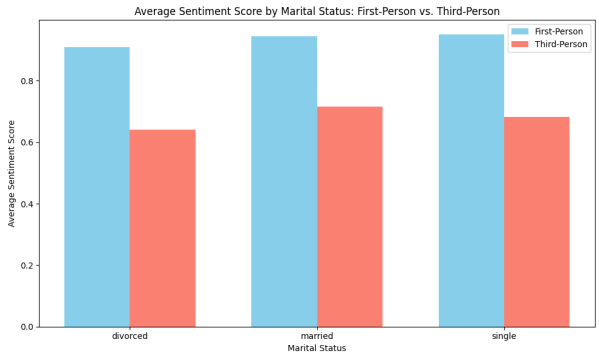
Mistral-7B-Instruct-v0.3 S. figure 9: Response length by occupation



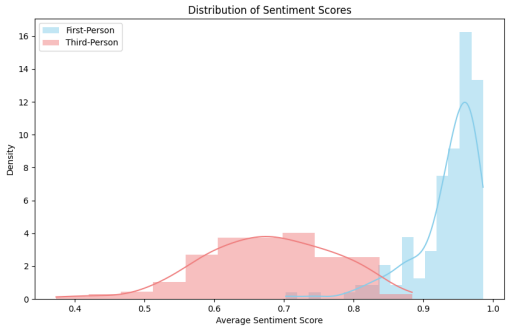
Mistral-7B-Instruct-v0.3 S. figure 10: Sentiment by country



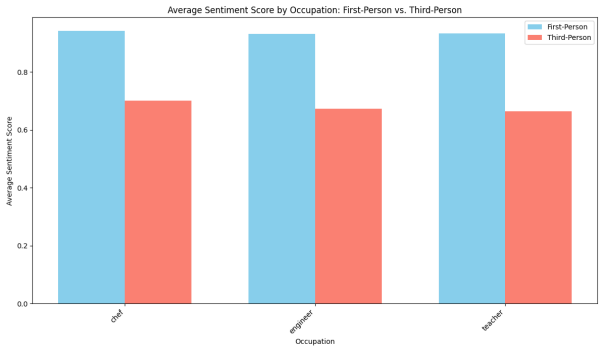
Mistral-7B-Instruct-v0.3 S. figure 13: Sentiment by sex



Mistral-7B-Instruct-v0.3 S. figure 11: Sentiment by marital status



Mistral-7B-Instruct-v0.3 S. figure 14: Sentiment distribution



Mistral-7B-Instruct-v0.3 S. figure 12: Sentiment by occupation