# Interaction with Artificial Social Agents

## A thematic analysis of people's experiences

**Celal Karakoç**[1]
**Supervisor: Prof. Dr. ir. W.P. Brinkman**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 19, 2025

Name of the student: Celal Karakoç

Final project course: CSE3000 Research Project
Thesis committee: Prof. Dr. ir. W.P. Brinkman, Prof. Dr. ir. R.L. Lagendijk

## Abstract

The use of Artificial Social Agents (ASAs) is rapidly expanding across society. As these agents become more integrated into our interactions, understanding the user experience of them becomes increasingly necessary to ensure their design aligns with user needs, promotes trust, and supports meaningful engagement. This study aims to investigate how users experience interactions with ASAs, focusing on using thematic analysis to identify recurring themes in user-reported experiences with ASAs. In addition, it also explores the reliability of locally hosted Large Language Models (LLMs) in identifying those experiences. We conducted a manual -peer validated- thematic analysis, resulting in a total of 31 themes. Afterwards, we conducted two experiments with LLMs, namely giving hem an unguided prompt (i.e. the LLM discovers and groups themes independently) and a guided prompt (i.e. the LLM matches predefined themes to responses) and measured their agreements with the manual analysis both intuitively and analytically. From our findings, it became clear that users experience ASAs through a balance of practical utility and emotional engagement. Themes covering the agent's helpfulness, sociability, enjoyability and perceived intelligence played a central role in shaping user experience. Most users responded positively to ASAs that felt intuitive, responsive, and human-like, though perceptions of human-likeness varied, sometimes enhancing the experience and other times creating discomfort. Our evaluation of LLMs showed that while they are capable of uncovering broad thematic patterns through unguided analysis, they fall short when tasked with consistently identifying and labeling predefined themes at the individual response level. This suggests that current LLMs, while useful as supplementary tools, are not yet reliable replacements for human-led thematic analysis in capturing the full nuance of user experiences at a detailed level. The conclusions reinforce the continued value and need of human-led thematic analysis, particularly when aiming to capture subtle, context-dependent insights that automated models may overlook.


## 1 Introduction

Artificial Social Agents (ASAs) are playing an increasingly prominent role in our everyday life as technologies and with it artificial intelligence continues to advance. From chatbots to voice assistants and robot vacuum cleaners, ASAs are envisioned as a step toward resolving issues of accessibility, emotional support, and user engagement across various domains. This development, however, brings with it its own set of challenges as not only functionality, but increasingly the *quality* of the social experience becomes more important. To facilitate this quality in Human-Computer Interaction with these social agents, users must be able to express their interest, wishes or queries in a natural and intuitive way, such as through speaking, typing, or gesturing [1, 2]. Yet despite growing interest in social interaction design, we lack a comprehensive understanding of how users subjectively experience their interactions with ASAs in natural, real-world contexts. For example, Fitrianie et al. [3] found that among 89 questionnaires reported in 81 papers from empirical studies reported in the intelligent virtual agent conference proceedings, the vast majority (over 76%) were used in only a single study and rarely reused, highlighting the lack of a clear consensus on which agent qualities users consistently value, critique, or find most meaningful. One promising avenue for exploring these perceptions is through the analysis of self-reported user experiences of user questionnaires. These narratives offer rich qualitative data that can be analyzed in a structured way which then can reveal patterns in how people engage with and evaluate ASAs. By examining these insights, we can better understand common themes in user experiences.

This feedback serves as a foundation for identifying recurring patterns and making sense of the diverse ways in which people engage with ASAs. To systematically explore these patterns, thematic analysis offers a flexible qualitative method for uncovering themes across user narratives, making it especially suited for capturing the complexity and nuance of subjective experience. Thematic analysis is particularly suited for this study as it provides a flexible approach to systematically identify, analyze, and report patterns of similarity and difference within qualitative data, allowing for the exploration of both anticipated and emergent themes in user experiences with ASAs, while still remaining flexible to unexpected findings [4]. Unlike more quantitative analysis approaches, thematic analysis provides the flexibility to identify both anticipated and unexpected patterns in user perceptions and it is not theoretically bounded [4]. This makes it ideal for analyzing open-ended questionnaire responses where the goal is to uncover recurring themes in how people evaluate and make sense of their interactions with ASAs.

However, this approach is not without its challenges. One such challenge is its recursive nature [5], which ensures that researchers iteratively move between different phases of the analysis. This ensures it being a time-consuming process. Large Language Models (LLMs) offer a potential solution for this issue. That said, it remains unclear whether LLMs can effectively support or replicate nuanced, human-led thematic analysis of such experiences [6, 7, 8].

This study explores **how people experience their interaction with Artificial Social Agents ($RQ_1$)**. Specifically, we examined which insights can be identified in user-reported experiences with Artificial Social Agents and which qualities users highlight when reflecting on these agents. To capture and evaluate these human-agent experiences, we analyzed a dataset collected through the Artificial Social Agent Questionnaire (ASAQ) [9], comprising of user-reported experiences with social agents. We then employed thematic analysis on these experiences to identify recurring themes and insights. Furthermore, we looked whether **a (locally hosted) LLM can identify these experiences ($SQ_1$)**, automating the time-intensive task of manual thematic analysis by delegating it to an LLM. After conducting both manual coding and analysis with a locally hosted LLM, we pose the question **how the manual and LLM-based thematic analyses compare ($SQ_2$)**.

# 2 Methodology

## 2.1 The dataset explained

The data used throughout this study is collected as part of a study into concurrent validation of ASAQ [10], wherein a normative dataset has been created as a community effort to develop a validated evaluation questionnaire. The ASAQs main purpose is trying to measure a standardized way of user experiences [11]. The data consists of two parts. Firstly, a dataset consisting of an open answer towards experiences with an agent. The participants were asked "*Please describe your experience with $agent$ in your own words (use at least 10 words, more words are welcome)*". There were a total of 666 responses to this question. These qualitative responses serve as the primary input for our thematic analysis. Secondly, a user study consisting of the 90 items from the long version of the ASAQ. Participants rated each item on a seven-point Likert scale [12] ranging from $-3$ (strong disagreement) to 3 (strong agreement), with 0 indicating neutrality. This scale reflects the degree to which participants agreed or disagreed with each construct. Participants will be denoted as *p#* with *#* being a number representing the response row. The items, which represent predefined constructs or dimensions of user experience [9, 10], were treated as complementary to the themes identified in the qualitative data. Finally, the dataset includes anonymized demographic descriptors such as age group, sex, education level, and geographic region. These variables served as contextual factors for interpreting both qualitative and quantitative findings.

## 2.2 Thematic Analysis

Thematic Analysis is a foundational method in qualitative research, employed to identify, analyze, and interpret patterns within data. It can be somewhat summarized by the six-phase process, namely familiarizing yourself with the data - generating codes - searching, reviewing, defining and naming themes - and finally producing the report [4]. There is an ongoing debate in thematic analysis which centers on whether it should adopt a more structured, descriptive approach or embrace a more interpretive reflexive methodology [13, 14]. This discussion is particularly pertinent in qualitative research, where the balance between methodological rigor and interpretive depth is crucial [15]. We decided to not fall towards either extreme, but take an approach more towards the middle (see Figure 1). We have conducted our manual thematic analysis by having a first pass throughout the data, familiarizing ourselves with the data, generating keywords from the responses. On our second pass, we generated codes from those keywords. While the generated codes were based on the keywords, a continuous reiteration of the responses, when codes were found through interpretation, was conducted, allowing for the emergence of new insights through a more intuitive analysis. This means in essence that our "passes" throughout our dataset are not strictly linear, but we continuously go back to the beginning of the data. These iterative, back-and-forth passes allowed us to revisit and refine emerging themes, resulting in multiple thorough reviews of the dataset. The codes were then summarized and abstracted on our "third pass" as themes found within the dataset. Inbetween the second and third pass a mind map was made, showing the relation between the codes and associated themes (see Figure 3), and it was updated throughout the other passes.

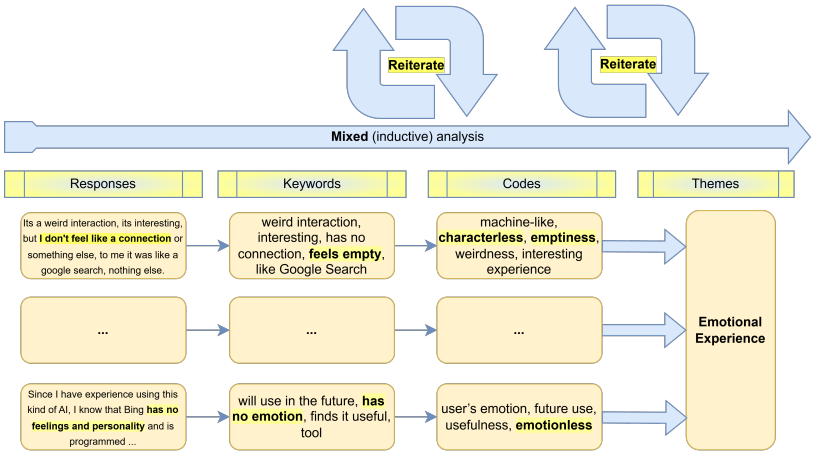


Figure 1: Thematic analysis manual approach.

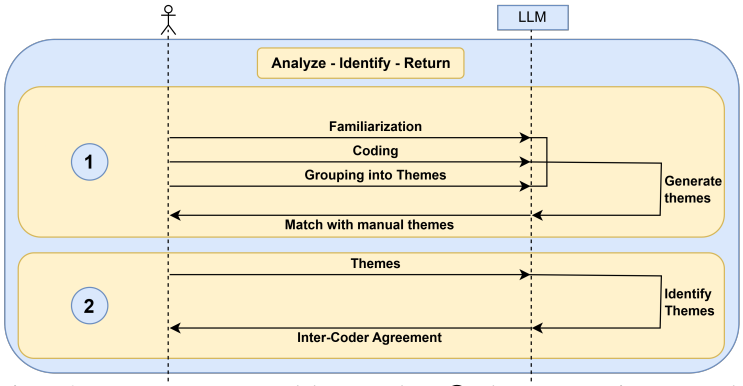


Figure 2: Large Language Model approaches: ① Theme generation approach. ② Theme application approach.

Afterwards, we are left with a set of themes and the conclusion of our manual thematic analysis. Following this, we tried to find out if an LLM can do our task of thematic analysis instead. For this we have tried out two approaches to test out the reliability of a generic LLM. As seen in Figure 2, this has been done to understand how well it formulates themes and how well it applies themes set to text. By breaking them up in this way we ensure that we can asses them easier without one biasing the other. To ensure that we do not specifically construct our prompt to get our specific results we generalized the prompt and gave it unchanged to a variety of LLMs. Furthermore, our prompts were constructed following an *Analyze - Identify - Return* approach, designed to broadly align with the six-phase thematic analysis process. This approach was inspired by the methodology used by Drápal et al. [16], though our implementation differs in specific details. For the first part, the LLM

was given the responses to find themes within them. So, we gave the LLM the prompt of a three-step process, asking it to Familiarize with the responses (*Analyze*), give a coding scheme (*Identify*) and group the codes together into coherent themes based on conceptual similarity or relation (*Return*). This approach does not specifically engineer the prompt to get the exact themes that we want in this occasion, but ensures a general reliability and the results to be broadly applicable to LLMs in general. This analysis was conducted in two runs, processing the first half of the data followed by the second half, after which the results were mapped to the manually identified themes. Afterwards, as our second approach, we tried to see how reliable the LLM would be by targeting the responses in unison rather than the text as a whole. Thus we gave it our themes from the manual analysis to analyze, and asked it to identify the themes on an individual basis and return per response the themes found within it. This then was compared to the results of the manual thematic analysis.

## 2.3 Reducing biases

The process as described by Figure 1 ensures that we can have the reproducibility of a more conservative approach of thematic analysis, while still maintaining flexibility and allowing interpretations. This flexibility does come at a cost, though, namely our own bias skewing our data. To minimize our bias in generating the themes, a peer researcher independently conducted the thematic analysis using their preferred method. We then assessed the consistency of our results through an Inter-Coder Agreement. The peer has been given the first $n = 100$ responses of our dataset to manually analyze. The themes have then been discussed with the peer and mapped intuitively towards the original resulting themes and any other remaining themes are either discarded if not meaningful or added towards the final themes. To ensure and unbiased comparison, we deliberately chose not to interfere with the peer's approach to thematic analysis, allowing them full autonomy in their method without imposing our own framework or biases. Accordingly, themes were only retained for analysis if they appeared more than three times ($N > 3$) in the peer's coding. Themes were also grouped by polarity, with positive and negative expressions, such as *Engagement* and *Lack of Engagement*, combined under broader thematic categories and considered as a single theme for the purpose of counting. Afterwards, the mind map (see Figure 3) is updated into a final state. In the same way as the first, a second Inter-Coder Agreement is held, but with the LLM as the peer. An intuitive mapping is conducted on the first prompt, and the second prompt is quantitatively analyzed with the manual thematic analysis being *the ground truth* and the LLM compared to it as its peer.

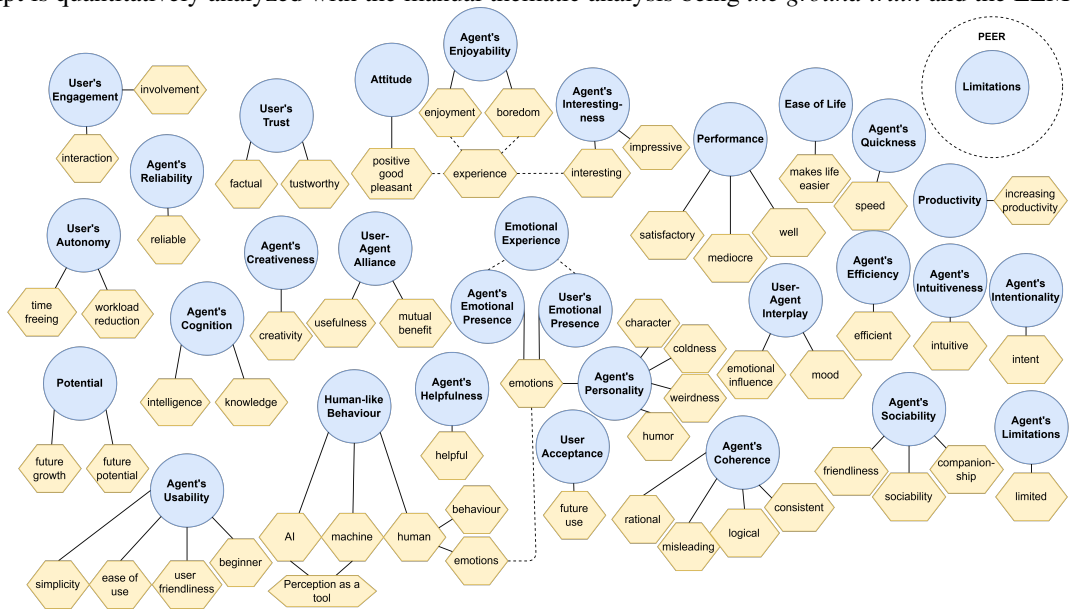

Figure 3: Mind map with regards to the codes (in yellow) and the resulting themes (in blue).

## 2.4 Quantitative Analysis

After getting the final set of themes, several quantitative analyses are conducted. To ensure that the Inter-Coder Agreement abstractly covers the same set of themes, besides the agreement itself, Cohen's Kappa Formula [17] is used, as seen in Appendix G. Table 1 shows the interpretation of our values and comparison between our themes with our peers' and also from our themes with the Large Language Models', showing whether LLMs can actually identify the same set of themes. Furthermore, a quantitative analysis of the themes in regards to the descriptors has been conducted to further substantiate the conclusions made and to see the reliability of the dataset itself.

Table 1: Interpretation of Cohen's Kappa values [18]

| Kappa ($\kappa$) | Interpretation |
|---|---|
| $< 0$ | Poor agreement |
| $0.00 - 0.20$ | Slight agreement |
| $0.20 - 0.40$ | Fair agreement |
| $0.40 - 0.60$ | Moderate agreement |
| $0.60 - 0.80$ | Substantial agreement |
| $0.80 - 1.00$ | Almost perfect agreement |

Table 2: Mapping of the themes with the ASAQ. Included is the second Inter-Coder Agreement, the Overlap coefficient (Szymkiewicz–Simpson coefficient [19]) calculated as: $Overlap(A, B) = \frac{|A \cap B|}{min(|A|,|B|)}$ and the overal agreement calculated by: $\frac{1}{n} \sum_{i=1}^{n} \text{Overlap}_i$.

| Theme | Mapping ASAQ (Coder 1) | Mapping ASAQ (Coder 2) | Overlap w. Coders (%) |
|---|---|---|---|
| **Agent's Cognition** | Coherence$^{(AC1)}$<br>Intentionality$^{(AI3)}$ | | 0 |
| **Agent's Coherence** | Coherence$^{(AC2\text{-}3)}$ | Coherence$^{(AC1\text{-}4)}$ | 100 |
| **Agent's Creativeness** | | | 100 |
| **Agent's Efficiency** | | | 100 |
| **Agent's Emotional Presence** | Emotional Experience$^{(AEI1\text{-}5)}$ | Emotional Experience$^{(AEI1\text{-}5)}$ | 100 |
| **Agent's Enjoyability** | Likeability$^{(AL2\text{-}3)}$<br>Emotional Experience$^{(AEI1, AEI3)}$<br>Enjoyability$^{(AE3\text{-}4)}$ | Likeability$^{(AL2\text{-}3)}$<br>Enjoyability$^{(AE1, AE3\text{-}4)}$ | 80 |
| **Agent's Helpfulness** | User-Agent Alliance$^{(UAL4)}$ | | 0 |
| **Agent's Intentionality** | Intentionality$^{(AI1\text{-}4)}$ | Intentionality$^{(AI1\text{-}4)}$ | 100 |
| **Agent's Interestingness** | Enjoyability$^{(AE2)}$ | Enjoyability$^{(AE2)}$ | 100 |
| **Agent's Intuitiveness** | Usability$^{(AU1\text{-}2)}$ | | 0 |
| **Agent's Limitation** | | | 100 |
| **Agent's Personality** | Personality Presence$^{(APP1\text{-}2)}$ | Personality Presence$^{(APP1\text{-}3)}$ | 100 |
| **Agent's Quickness** | Usability$^{(AU3)}$ | | 0 |
| **Agent's Reliability** | User Trust$^{(UT3)}$ | User Trust$^{(UT3)}$ | 100 |
| **Agent's Sociability** | Sociability$^{(AS1\text{-}3)}$ | Sociability$^{(AS1\text{-}3)}$ | 100 |
| **Agent's Usability** | Usability$^{(AU1\text{-}2)}$ | Usability$^{(AU1\text{-}3)}$ | 100 |
| **Attitude** | Enjoyability$^{(AE4)}$<br>User Attitude$^{(AT1\text{-}3)}$ | User Attitude$^{(AT1\text{-}3)}$ | 100 |
| **Ease of Life** | | | 100 |
| **Emotional Experience** | Emotional Experience$^{(AEI\text{-}5, UEP1\text{-}4)}$<br>Likeability$^{(AL5)}$ | Emotional Experience$^{(UEP1\text{-}4)}$<br>User-Agent Alliance$^{(UAL6)}$ | 80 |
| **Human-like Behaviour** | Believability$^{(HLB1\text{-}5, NB2)}$ | Believability$^{(HLB1, HLB3\text{-}4, NB2\text{-}3)}$ | 80 |
| **Performance** | Performance$^{(PF1)}$ | Performance$^{(PF1\text{-}3)}$ | 100 |
| **Potential** | | | 100 |
| **Productivity** | | | 100 |
| **User Acceptance** | User Acceptance$^{(UAA1\text{-}3)}$ | User Acceptance$^{(UAA1\text{-}3)}$ | 100 |
| **User's Autonomy** | | | 100 |
| **User's Emotional Presence** | Likeability$^{(AL5)}$ | | 0 |
| **User's Engagement** | User Engagement$^{(UE1\text{-}3)}$ | User Engagement$^{(UE1\text{-}3)}$ | 100 |
| **User's Trust** | User Trust$^{(UT1\text{-}2)}$ | User Trust$^{(UT1\text{-}3)}$ | 100 |
| **User-Agent Alliance** | Likeability$^{(AL4\text{-}5)}$<br>User-Agent Alliance$^{(UAL1\text{-}3)}$ | Likeability$^{(AL4)}$<br>User-Agent Alliance$^{(UAL1\text{-}5)}$ | 80 |
| **User-Agent Interplay** | Emotional Experience$^{(UEP2)}$<br>User-Agent Interplay$^{(UAI1, UAI4)}$ | User-Agent Interplay$^{(UAI1\text{-}4)}$ | 66 |
| **Limitations** | | | 100 |
| **Overall Agreement** | | | 80.19 |

*Note:* The construct and items of the ASAQ mapping with their definitions can be found at the questionnaire [10] or website [20]. The themes were mapped on the item-level and the long version of the ASAQ was used. The items are put as subscript above the constructs they are under. A dash (-) between indicates *a* through *z*, i.e. *AT1-3* indicates *AT1, AT2, AT3*.

### 2.5 Correlation with regards to the ASAQ

The dataset of the ASAQ also consisted of a 7-point ranking from the 90 items of the ASAQ, as discussed in the introduction. To see whether the responses to the open question and the ratings correlated, we created a mapping with our themes and the

predefined constructs of the ASAQ. Afterwards, the entire dataset is correlated with the mapped constructs. Since the dataset consists of ordinal data, Spearman's Rank Correlation ($\rho$) is used (as seen in Appendix G). The interpretation of Spearman's correlation, in terms of its significance and strength can vary, but for our dataset to interpret the correlations with the ASAQ we applied the classification system used in Psychology as described by Dancey and Reidy [21, 22]. A fourth pass through the dataset has been conducted, noting the direction of a theme in either positive (1), neutral (0) or negative ($-1$). As an example, take the theme *User's Trust*. If the participant indicates that the agent was trustworthy, the theme is deemed positive, so a value of 1 is assigned. In the same manner, if the participant indicates that the agent was untrustworthy a value of $-1$ is assigned. If the participant deemed the agent, neither trustworthy nor untrustworthy, a value of 0 is assigned. If the theme was not identified, no value is assigned and the theme is ignored in its entirety. This then is correlated with the mapped constructs for said theme from the ASAQ, with values between $-3$ and 3. What themes are mapped to which constructs from the ASAQ can be seen in Table 2. To address the potential bias in the mapping, a peer independently performed the mapping of themes to the ASAQ. An Inter-Coder Agreement was then conducted to establish a final mapping.

## 3 Results

Before going to the final themes, we begin addressing the comparison and resolvement of our discrepancies and the results of our Inter-Coder Agreements in Tables 2 and 3. The process conducted by the peer can be seen in detail in Appendix C, with Figure 8 the approach used and with Tables 9 and 10 the definitions and mapping respectively. In regards to the first Inter-Coder Agreement, the themes derived from our peer, held up with Cohen's Kappa as seen in Table 3, with a mean $\kappa$ of $0.64$, indicating a substantial agreement. There was generally a high level of agreement, with the **a** and **d** values being high, indicating consistent identification of themes as present or absent, respectively. In cases where $\kappa$ fell below the threshold for moderate agreement ($\leq 0.6$), the **b** value, indicating that the theme was marked present by Coder 1 but not by Coder 2, notably increased, as can seen in Appendix C Table 11. In this way, even the low agreement values can be explained away, with Coder 2 being the peer, since our approach was reiterative by nature, which resulted in finding a lot more themes as we iterated through with multiple reiterative passes.

Table 3: Comparison of themes with those derived by a peer (Coder 2), based on a sample of $n = 100$ responses.

| **Theme (Coder 1)** | **Theme (Coder 2)** | $\kappa$ | Interpretation $\kappa$ |
|---|---|---|---|
| Agent's Coherence | Accuracy | 0.83 | Almost perfect agreement |
| Agent's Creativeness | Creativity | 0.92 | Almost perfect agreement |
| Agent's Efficiency | Efficiency | 0.93 | Almost perfect agreement |
| Agent's Enjoyability | Enjoyability | 0.71 | Substantial agreement |
| Agent's Helpfulness | Helpfulness | 0.79 | Substantial agreement |
| Agent's Interestingness | Interestingness | 0.28 | Fair agreement |
| Agent's Usability | Usability, Accessibility, Convenience | 0.8 | Almost perfect agreement |
| Attitude | Entertainment | 0.2 | Fair agreement |
| Emotional Experience | Emotional Connection | 0.33 | Fair agreement |
| Human-like Behaviour | Human-like Behavior | 0.5 | Moderate agreement |
| Potential | Potential | 0.65 | Substantial agreement |
| Productivity | Productivity | 0.74 | Substantial agreement |
| User's Engagement | Engagement | 0.63 | Substantial agreement |
| User's Trust | Trust | 0.71 | Substantial agreement |

In regards to the second Inter-Coder Agreement, as seen in Table 2, which focused on the mapping to the ASAQ, the peer mapped the final set of themes onto the 90 items of the ASAQ, where applicable. The overall overlap between the themes mapped was 80.19%. This high degree of overlap suggests a strong consistency in how themes were associated with the constructs, suggesting that the mapping process may have been relatively reliable and less influenced by subjective bias. Consequently, the results supported the robustness of the thematic alignment within the ASAQ framework. The strong agreement observed here enhances confidence in the validity of the mapping methodology and implies that the identified constructs effectively capture the underlying themes. This ensures that subsequent analyses and interpretations that rely on these mappings are grounded in stable and reproducible foundations.

Our final themes and their definitions can be seen in Table 4, with quotes present in Appendix B Table 8 to provide a more comprehensive understanding. *Limitations* is the only theme from the peer added towards our final themes after the first Inter-Coder Agreement, as also shown in our mind map in Figure 3. From our analysis, as shown in Figure 4, the themes throughout were quite evenly divided from their descriptors. There were no observed outliers and as the count of the themes increased, the perceived divisions also seem to become non-existent. With 254 *Emotional Experience* was by far the most occurring theme (see Appendix D Table 12), which is why two subsets *User's Emotional Presence* and *Agent's Emotional Presence* were created to subdivide the theme making it less abstract and more specific. The themes that mattered the most towards the experience of people as per the counts were the *Agent's Helpfulness*, *Attitude* and *Human-like Behaviour*. The prominence of these themes

Table 4: Final Themes and Definitions with example quotes found in Appendix B Table 8

| Theme | Abbreviation | Definition |
|---|---|---|
| **Agent's Cognition** | COG | The agent is intelligent/knowledgeable. |
| **Agent's Coherence** | COH | The agent is perceived as logical and consistent. |
| **Agent's Creativeness** | CRE | The agent is perceived as creative. |
| **Agent's Efficiency** | EFF | The agent is perceived as efficient. |
| **Agent's Emotional Presence** | EMP | The user's perception of the agent's emotions during and after interaction. |
| **Agent's Enjoyability** | ENJ | The extent to which the user finds the interaction with the agent enjoyable/boring. |
| **Agent's Helpfulness** | HLP | The agent is perceived as helpful. |
| **Agent's Intentionality** | INT | The agent is perceived as acting deliberately and with intention. |
| **Agent's Interestingness** | INS | The extent to which the user finds interaction with the agent interesting. |
| **Agent's Intuitiveness** | ITU | The extent to which the agent is perceived as intuitive. |
| **Agent's Limitation** | LIM | The user perceives the agent as being useful only for limited use/purposes. |
| **Agent's Personality** | PER | The distinctive combination of character traits/qualities of the agent (or lack thereof). |
| **Agent's Quickness** | QCK | The extent to which the agent performs tasks quickly. |
| **Agent's Reliability** | REL | The agent is perceived as reliable. |
| **Agent's Sociability** | SOC | The user perceives the agent as sociable. |
| **Agent's Usability** | USA | The user perceives the agent as easy to use, user- or beginner-friendly, or simple to interact with. |
| **Attitude** | ATT | The extent to which the user finds the interaction with the agent positive. |
| **Ease of Life** | EOL | The agent is perceived as making the user's life easier. |
| **Emotional Experience** | EMX | A self-contained emotional experience during interaction. |
| **Human-like Behaviour** | HLB | The agent behaves like a human, expressively or emotionally, or conversely, like a machine/AI/tool. |
| **Limitations** | LIP | User thoughts on things it cannot do well or problems/limitations noticed |
| **Performance** | PRF | The extent to which the agent performs tasks well. |
| **Potential** | POT | The user perceives the agent having future potential for improvement. |
| **Productivity** | PRO | The agent helps increase the user's productivity. |
| **User Acceptance** | UAC | The likelihood that the user will use the agent again or in the future. |
| **User-Agent Alliance** | UAA | The extent to which the user and agent collaborate for mutual benefit. |
| **User-Agent Interplay** | UAI | The degree to which the user and agent influence each other. |
| **User's Autonomy** | AUT | The user perceives the agent reducing the user's workload and allowing for more free time. |
| **User's Emotional Presence** | UEP | The user's emotional state during and after interacting with the agent. |
| **User's Engagement** | ENG | The extent to which the user feels involved in the interaction with the agent. |
| **User's Trust** | TRU | The user perceives the agent as trustworthy and factual. |

suggests participants prioritize emotional connection and practical utility in interactions with ASAs. In regards to *Human-like Behaviour* the agents were mostly seen and perceived as a tool that tries to act like a human ("Alexa is a tool for certain actions that can be automated and don't need my full attention" *p87*), but sometimes also as a companion ("Alexa is a friend that we need in serious times." *p439*). As also noticeable in the first quote, this did not substantially alter the participants' evaluations of the agent's usability, helpfulness, coherence or other such factors. Furthermore, some users appreciated anthropomorphic traits and found them comforting or engaging ("...the voice i enjoyable : )" *p421*, "i never seen such rudeness from a bot, so it was quite fun and interesting" *p643*), while others were put off by them ("...the voice usually reminds you that it is a program and it affects the interaction." *p22*). *Agent's Cognition* and *Agent's Coherence* received moderate counts and have a mixed polarity, suggesting participants noticed both strengths and weaknesses in the agent's intelligence and logical consistency, which reflects a nuanced perception of the agents' cognitive abilities. Participants may have appreciated moments when the agent demonstrated smart or insightful responses ("I was impressed with ChatGPT. It engages you in a conversation and can also refer to what was said before." *p274*), yet were also sensitive to lapses in coherence, such as contradictory, off-topic or downright rude replies ("...I was stuck because she want me to explain 1 sentence when i was asking which one." *p666*). This mixed sentiment indicates that while users recognize the potential of the agent's cognitive functions, inconsistencies in reasoning or dialogue flow can diminish trust and overall satisfaction. These results highlight the importance of both depth of knowledge and conversational stability in shaping user experience with Artificial Social Agents.

A portion of the themes can be broadly grouped based on the participant's experience of the agent's functional and affective characteristics. In terms of functional characteristics, we find that participants evaluated the *Agent's Reliability*, *Quickness*, and *Efficiency* based on concrete outcomes or performance-related experiences ("Copilot provided autocomplete-style suggestions as I coded. It made it easier and faster to write code more efficiently." *p73*). These were generally derived from tasks in which the agent either succeeded or failed in delivering responses efficiently and accurately. These attributes were easier to assess

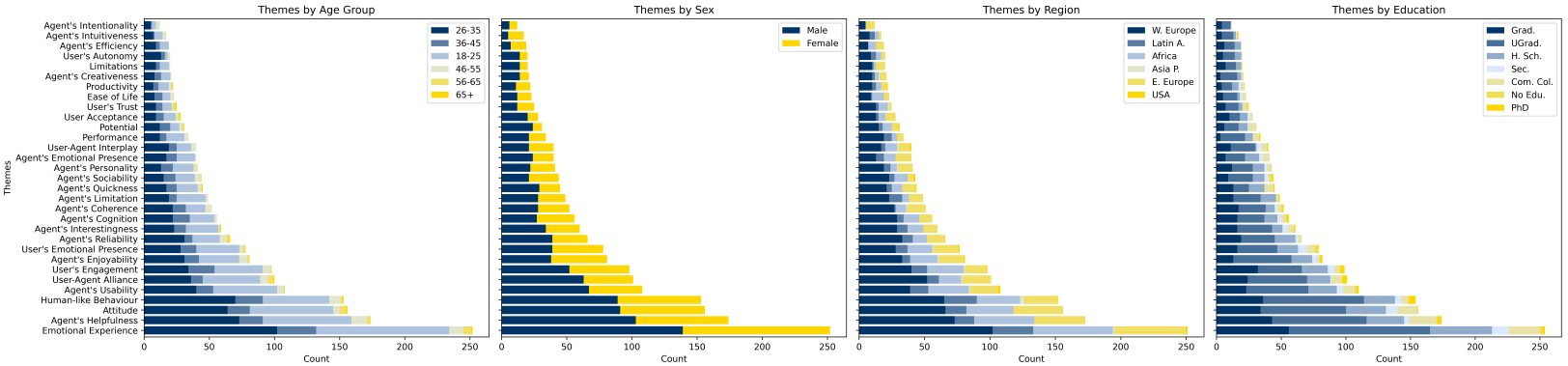


Figure 4: All the themes that were found divided into their respective descriptors sorted by their counts. In more detail in Appendix D Figures 9, 10, 11, 12

for participants, and consequently, these themes appeared consistently across various agent interactions. In terms of *Usability*, the responses mostly looked at it through the lens of the user- or beginner-friendliness of the agent and how simple or easy to use the agent was ("I find it very easy to work with" *p404*, "using bard is easy, its interface is friendly" *p479*, "It was a simple to use AI tool" *p433*). Conversely, affective characteristics, such as *Agent's Sociability*, *Personality*, *User's Engagement*, and both *User's and Agent's Emotional Presence* capture more abstract elements of the interaction, relating to the user's perception of the agent's demeanor or presence rather than its capabilities. The User's own personal emotions also played a huge role in the perceivement of these attributes. A more cold tone in language (e.g. "the bot was not . . . to go nowhere" *p608*) as opposed to a warmer tone (e.g. "Its the future! I think everyone needs a little more Bard in their lives. Amazing tool and comforting companion." *p45*) resulted in other themes also rated more negatively or positively, respectively.

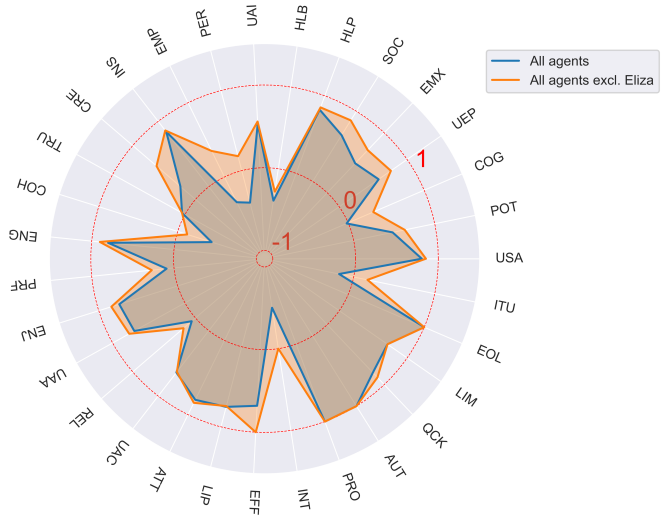


Figure 5: Average polarity of participant themes with 1 being positive and -1 being negative labeled manually within all 666 user responses.

Finally, it is worth highlighting that the agents were overwhelmingly evaluated positively, with *Eliza*, an early rule-based chatbot simulating a psychotherapist [23], standing out as an exception. Participants often described *Eliza* as "strange", "not enjoyable" or "unpleasant". Even in cases where *Eliza* was described as interesting, this perception was often accompanied by a negative evaluation of the interaction overall (e.g. "It was an interesting experience. I would not say it was a pleasant interaction." *p587*). This is noticeable in Figure 5, which shows that the average direction from all themes move towards positive when excluding *Eliza* from the equation. The only exception seem to be *Human-Like Behaviour*. The participants seem to think of the social agents more as tools to be used rather than companions or friends that can be talked to.

Table 5: Correlations with the ASAQ.

| Theme | $\rho$ | $p$ -value | CI (95%) | Correlation | Significance |
|---|---|---|---|---|---|
| Agent's Cognition | 0.56 | <0.001 | [0.36, 0.72] | Moderate positive correlation | Very strong statistical significance |
| Agent's Coherence | 0.41 | 0.003 | [0.15, 0.61] | Moderate positive correlation | Strong statistical significance |
| Agent's Emotional Presence | 0.23 | 0.148 | [-0.08, 0.5] | Weak positive correlation | No statistical significance |
| Agent's Enjoyability | 0.59 | <0.001 | [0.43, 0.72] | Moderate positive correlation | Very strong statistical significance |
| Agent's Helpfulness | 0.12 | 0.123 | [-0.03, 0.26] | Weak positive correlation | No statistical significance |
| Agent's Intentionality | 0.5 | 0.096 | [-0.1, 0.84] | Moderate positive correlation | No statistical significance |
| Agent's Interestingness | 0.32 | 0.011 | [0.08, 0.53] | Moderate positive correlation | Statistically significant |
| Agent's Intuitiveness | 0.59 | 0.013 | [0.15, 0.83] | Moderate positive correlation | Statistically significant |
| Agent's Personality | 0.3 | 0.053 | [0, 0.55] | Moderate positive correlation | No statistical significance |
| Agent's Quickness | 0.11 | 0.485 | [-0.19, 0.39] | Weak positive correlation | No statistical significance |
| Agent's Reliability | 0.53 | <0.001 | [0.32, 0.68] | Moderate positive correlation | Very strong statistical significance |
| Agent's Sociability | 0.63 | <0.001 | [0.41, 0.78] | Strong positive correlation | Very strong statistical significance |
| Agent's Usability | 0.25 | 0.008 | [0.07, 0.42] | Weak positive correlation | Strong statistical significance |
| Attitude | 0.25 | 0.002 | [0.1, 0.39] | Weak positive correlation | Strong statistical significance |
| Emotional Experience | 0.32 | <0.001 | [0.21, 0.43] | Moderate positive correlation | Very strong statistical significance |
| Human-like Behaviour | 0.37 | <0.001 | [0.23, 0.5] | Moderate positive correlation | Very strong statistical significance |
| Performance | 0.33 | 0.055 | [-0.01, 0.6] | Moderate positive correlation | No statistical significance |
| User Acceptance | 0.3 | 0.115 | [-0.08, 0.61] | Moderate positive correlation | No statistical significance |
| User's Emotional Presence | 0.45 | <0.001 | [0.25, 0.61] | Moderate positive correlation | Very strong statistical significance |
| User's Engagement | -0.05 | 0.594 | [-0.25, 0.14] | Weak negative correlation | No statistical significance |
| User's Trust | 0.47 | 0.018 | [0.09, 0.73] | Moderate positive correlation | Statistically significant |
| User-Agent Alliance | 0.26 | 0.009 | [0.07, 0.43] | Weak positive correlation | Strong statistical significance |
| User-Agent Interplay | 0.53 | <0.001 | [0.26, 0.72] | Moderate positive correlation | Very strong statistical significance |

● Strong ($|\rho| \geq 0.5$) ● Moderate ($0.3 \leq |\rho| < 0.5$) ● Weak ($0.1 \leq |\rho| < 0.3$) ● Negligible ($|\rho| < 0.1$)

In examining the correlations derived from the mapping with the ASAQ (see Table 5), all themes that showed statistical significance exhibited a positive correlation with their ASAQ counterparts. The confidence intervals of these themes also fall entirely within the positive range. This indicates that the more positively users rated a theme (e.g. how smart, enjoyable or reliable the agent seemed), the higher their overall ASAQ score tended to be. Most of the themes demonstrate a moderate positive correlation, which supports the overall conclusion, although in several cases the confidence intervals lean toward suggesting a stronger correlation can be derived. Notably, the *Agent's Cognition*, *Enjoyability*, *Reliability*, and *Sociability* themes show particularly strong correlations with ASAQ, suggesting that users place high value on these aspects. Additionally, *Emotional Experience* and *User's Emotional Presence* were found to be positively correlated with a strong statistical significance. Given the conceptual closeness between *User's* and *Agent's Emotional Presence*, both being a subset *Emotional Experience*, we expected *Agent's Emotional Presence* to show a similar pattern. However, this was not supported by the data. Also, it is important to note that our cutoff points should be used judiciously, as explained by Schober et al. [24], as they should be used in accordance with their strength in the context of the research question.

Interestingly, although several themes show wide confidence intervals, they still fall entirely within the positive range. This suggests that while the magnitude of the correlation may be uncertain, the direction is consistently positive. For example, *Agent's Intuitiveness* has a relatively wide CI but still supports a meaningful positive correlation with the ASAQ. This pattern reinforces the idea that, even when variability exists, the underlying relationship remains directionally stable. Among all themes, *Agent's Sociability* stands out with the highest rho value ($\rho$ = 0.63), and a relatively narrow confidence interval. Similarly, high rho values in other themes like *Enjoyability* and *Reliability* support the conclusion that users deeply value personal and interpersonal characteristics in agent interactions. Furthermore, as the level of statistical significance increases (i.e., lower p-values), confidence intervals tend to narrow and remain consistently in the positive range, indicating that these results are unlikely to be due to chance.

The only two outliers in relation to our expectations are the themes *Agent's Quickness* and *User's Engagement*. We noticed that for *Agent's Quickness* the Inter-Coder Agreement of the peer did not map it towards an ASAQ theme, and there was already some uncertainty as to whether it fully corresponds to any ASAQ construct. This may suggest that our initial doubts were valid, or alternatively, the result may simply reflect the true nature of this theme. In terms of *User's Engagement*, both analyses were mapped towards the same ASAQ constructs, but it still has a weak negative correlation, which is quite an unexpected finding. However, the confidence interval indicates that this result is not conclusive, as the correlation could just as easily be positive. Any conclusions based on the *User's Engagement* theme should be interpreted with caution. Overall, the correlations support our previous findings, as we can derive our conclusions from two separate parts of the questionnaire.

### 3.1 Large Language Models

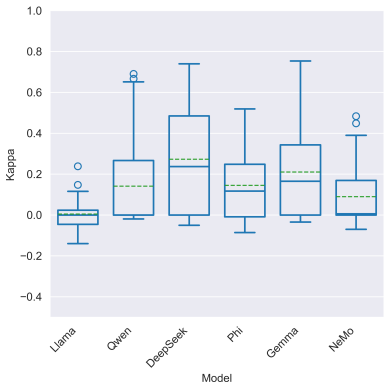


Figure 6: Kappa Distribution across LLMs

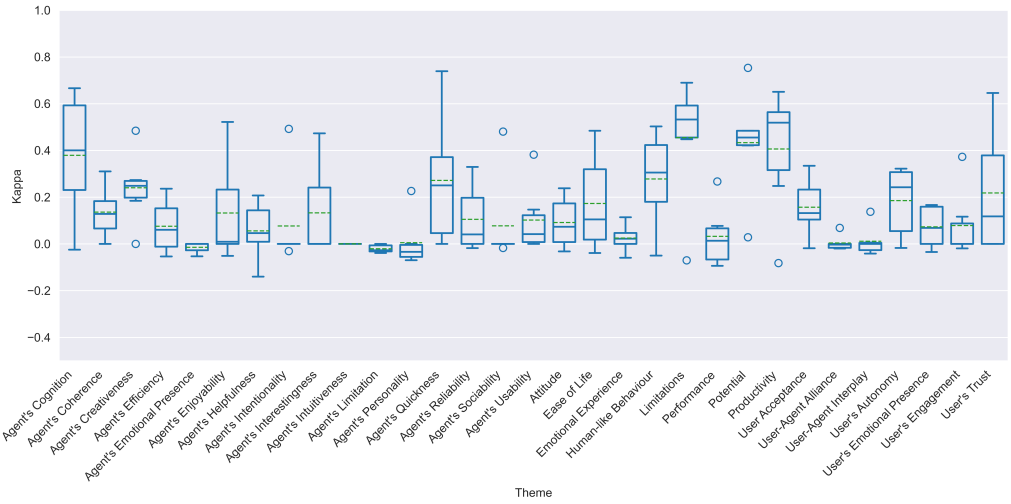


Figure 7: Kappa Distribution across Themes

The results indicate that the LLM performs poorly when applying the coding scheme through a guided prompt. As shown in Figure 6 and explicitly in Appendix E Table 13, the average $\kappa$ over all themes is very low. Additionally, while lowering the temperature can make the output a lot more deterministic, the outcomes remain largely unchanged regardless of the temperature setting. Theme application appears to be quite the challenge for an LLM. Even in cases where the interpretation accuracy seems somewhat high, it remains inconsistent across themes, and the themes predicted more accurately are not consistently identified by all LLMs at all and they are not consistently the same throughout LLMs. Unfortunately, as shown in Figure 6, the ability of Large Language Models to reliably apply predefined themes to a set of responses remains consistently low across models. Even DeepSeek, the best-performing model, does not have its upper quartile exceed the threshold for moderate agreement, while all other models fail to reach even that level of consistency. This pattern remains consistent when examining the theme-level performance in Figure 7. Notably, no theme's upper quartile exceeds the threshold for moderate agreement. This stands in contrast to the peer agreement values presented in Table 3, where all values at or below moderate agreement can be explained by the **b** value being so high, while the remaining themes demonstrate relatively strong agreement scores, with a substantial amount even reaching an almost perfect agreement interpretation. Even the four best-performing themes, *Agent's Cognition*, *Limitations*, *Potential* and *Productivity*, do not show upper quartiles surpassing moderate agreement. Moreover, *Agent's Cognition* and *Productivity* exhibit a notably wide inter-quartile range, indicating substantial variability in how consistently these themes are applied by the LLMs.

The unguided prompt approach, where LLMs generate themes based on the content of the text, yielded more promising results, as shown in Table 6. A notable pattern across models was their tendency to consolidate multiple specific themes into broader, unifying themes. Despite this generalization, the themes identified by the LLMs were indicated in the text, suggesting that the models were able to extract the underlying thematic structure of the responses with good accuracy. The few themes that were not identified by the models primarily concerned practicality and ethical concerns. In the case of practicality, the model-generated content referred to related ideas, such as usefulness, convenience or efficiency, but did not explicitly match the these aspects as it was in the original coding scheme. So we could not clearly align it with any of our themes. In regards to the unmapped themes regarding ethical concerns (i.e. *Ethical Considerations* and *Ethical Concerns*), these themes were absent not only in the LLM outputs but also in both manual codings. Although some indication of the theme was present in the peer's manual coding (*Cultural Limitations* to be precise), this was thrown out in the Inter-Code Agreement. This oversight suggests that more Inter-Coder Agreements with more coders may have helped identify this theme earlier, highlighting the value of more perspectives in qualitative analysis.

Table 6: Mapping of LLM themes ranked by frequency, high to low.

| Manual | Llama | Qwen | DeepSeek | Phi | Gemma | NeMo |
|---|---|---|---|---|---|---|
| **Emotional Experience** | Emotional Connection[2] | Emotional and Social Experience[2] | Emotional Engagement[1]<br>Emotional Engagement[2] | Interaction Quality[1] | | |
| **Agent's Helpfulness** | Helpfulness and Convenience[1]<br>Helpfulness[2] | | Practical Functionality[1]<br>Interaction Quality[2] | Functional Benefits[2] | Utility and Efficiency[1]<br>Utility and Efficiency[2] | |
| **Attitude** | | Emotional and Social Dimensions[1] | | | | |
| **Human-like Behaviour** | Natural Interaction[2] | Emotional and Social Dimensions[1] | Perceived Humanity[1]<br>Human vs. Machine Interaction[2] | Interaction Quality[1] | The Human Connection[1]<br>The Illusion of Humanity[2] | Human-Like Interactions and Limitations[1] |
| **Agent's Usability** | User Experience[1]<br>User Experience[2] | Usability and Accessibility[1]<br>Usability and Practicality[2] | Interaction Quality[2] | Usability[1]<br>User Experience[2] | Usability and Accessibility[1]<br>Utility and Efficiency[2] | Practical Assistance[1] |
| **User-Agent Alliance** | | | | | Utility and Efficiency[1] | Practical Benefits[2] |
| **User's Engagement** | Entertainment and Engagement[1] | | | | | |
| **Agent's Enjoyability** | Entertainment and Engagement[1] | Emotional and Social Experience[2] | | User Experience[2] | | |
| **User's Emotional Presence** | Emotional Connection[2] | Emotional and Social Experience[2] | Emotional Engagement[1]<br>Emotional Engagement[2] | | | |
| **Agent's Reliability** | Accuracy and Reliability[1]<br>Reliability and Trust[2] | | Trust and Reliability[1] | Performance[1]<br>Functional Benefits[2] | Trust and Reliability[1]<br>Utility and Efficiency[2] | |
| **Agent's Cognition** | | Perceived Intelligence vs. Limitations[1] | | | | |
| **Agent's Coherence** | | | | Performance[1] | | |
| **Agent's Limitation** | | Task-Specific Utility and Limitations[2] | | | | |
| **Agent's Quickness** | | | | | | Practical Assistance[1] |
| **Agent's Sociability** | Emotional Connection[2] | Emotional and Social Experience[2] | | | Social Connection[2] | |
| **Performance** | | Functional Utility and Efficiency[1] | Practical Functionality[1] | | | |
| **Potential** | | | Future Outlook[1] | Development Needs[1] | | Learning and Growth[1] |
| **User's Trust** | Reliability and Trust[2] | Emotional and Social Dimensions[1] | Trust and Reliability[1] | | Trust and Reliability[1] | |
| **Agent's Creativeness** | | Perceived Intelligence vs. Limitations[1] | | | | |
| **User's Autonomy** | | | | | | Practical Assistance[1] |
| **Limitations** | Limitations and Biases[1]<br>Limitations and Frustration[1] | Perceived Intelligence vs. Limitations[1] | Limitations and Challenges[1]<br>Frustration and Limitations[2] | Performance[1]<br>Challenges and Concerns[2] | User Frustration and Limitations[2] | Human-Like Interactions and Limitations[1]<br>Reliability Issues[2] |
| **Agent's Efficiency** | | Functional Utility and Efficiency[1] | | | Utility and Efficiency[1] | Practical Assistance[1] |
| **Agent's Intuitiveness** | | | | User Experience[2] | Usability and Accessibility[1] | |
| **Unmapped** | Practical Applications[1] | Ethical and Cultural Considerations[1]<br>Ethical and Privacy Concerns[2] | | Application[1]<br>Ethical Considerations[1] | Ethical Considerations and Future Concerns[1]<br>Unique Experience[2] | Emotional Intelligence Gap[2] |

*Note:* run 1 and run 2 are denoted by the superscripts. Definitions and details of the themes are available in the GitHub repository [25].

## 4 Discussion

This study set out to explore how users perceive and engage with Artificial Social Agents through thematic analysis of open-ended responses and to examine whether Large Language Models could support or replicate this process. This was done by trying to answer the following questions:

**$RQ_1$** *How do people experience their interaction with Artificial Social Agents?*
**$SQ_1$** *Can a (locally hosted) Large Language Model (LLM) identify these experiences?*
**$SQ_2$** *How do manual and LLM-based thematic analysis compare with each other?*

For the main question, people experience their interactions with ASAs as generally positive and interesting, often resulting with a desire to continue using the agent ("Using CoPilot was interesting, and I would definitely continue using it." *p341*). This is evident in the relatively high count and strong positive polarity of both *Attitude* and *Agent's Interestingness*, which reflects users' overall experience with their interactions. Participants typically perceived ASAs more as tools than as companions. This perspective was also apparent in responses where the agents did not behave in a human-like way ("It was easy to understand and navigate, it felt like a human experience just without the emotions." *p94*, ("it felt like speaking to a bot, i didn't feel any actual interaction between us" *p579*)) or where limitations for the agent were noticed ("My only issue with Google Assistant is that it doesn't always understand my accent. ... It's a great tool but needs to be modified to accommodate the whole world." *p9*, "We (humans) like to interact with machines as if they are living beings, ... But we're still aware of the machine's limitations." *p254*) both reflected in the themes *Human-like Behaviour* and *Limitations*.

The agents were appreciated for the value they offered in how they contributed to users' daily lives. This is supported by the notably high polarity scores for *Ease of Life*, *Productivity*, *Autonomy*, and *Helpfulness*, all of which highlight how ASAs helped users save time, complete tasks more efficiently, and feel more supported and autonomous ("ChatGPT has made my life so much easier and has contributed significantly in my writing." *p412*, "The Roomba ... it's a time-saving marvel!" *p487*). Furthermore, the perceived intelligence of the agent also played a critical role in shaping the user experience. However, it was a double-edged sword: once users detected incoherent responses, contradictions, or awkward behavior, their perception shifted negatively ("It was quite awkward. I felt like she didn't listen or understand me at all." *p665*). This shift is reflected in the neutral polarity of themes like *Agent's Cognition* and *Agent's Coherence*, indicating that these qualities could make or break the experience depending on how well the agent performed in context.

Participants highly appreciated agents that were useful in completing tasks either for a general purpose or a singular purpose ("Google Assistant has been very helpful and easy to use." *p93*, "I use Google assistant mostly to check the weather, and it always provides me with good information. " *p84*), as noticeable by *Agent's Helpfullness* and *Agents Limitation*. Another important theme was the user's perception of partnership with the agent. Many users appreciated a sense of collaboration and mutual benefit in their interactions ("It is comfortable, and it is versatile so I can do a lot of things with it, my brother is blind and it serves a lot of use for him in terms of accessibility." *p520*), which was reflected in the theme *User-Agent Alliance*.

The tone also shaped experiences. Friendly, warm tones enhanced user perceptions, while robotic or cold responses had a detrimental effect. Interestingly, enjoyable and entertaining interactions contributed to positive evaluations, even in the case of *Eliza*, where sometimes the agent's rudeness was perceived as entertaining and humorous, thereby creating a positive experience despite low practical value ("she was very rude and seems confused about every interaction, but could have asked me to explain things more nicer... i never seen such rudeness from a bot, so it was quite fun and interesting" *p643*, "I did not enjoy using Eliza at all. I found it very hard to read her tone and she came off very abrasive and accusatory." *p664*). Themes such as *Agent's Sociability* and *Human-like Behaviour* revealed a split in user preference: while some appreciated anthropomorphic traits, others were unsettled by them (" Some questions were asked and the answers were interesting and captivating. Socialization was always positive" *p17*, "It was an interesting experience. I tried other AI's and I can say that the voice usually reminds you that it is a program and it affects the interaction. However it did its job." *p22*).

As a final point, the themes developed through this thematic analysis illustrate the diverse ways in which participants make sense of their interactions with Artificial Social Agents. However, some themes, such as *Emotional Experience*, encompass broad reflections on the interaction, while others, like *Agent's Quickness*, are more narrow in scope, revealing a potential gap in how certain aspects of user experience are captured and suggesting the need for more coders conducting the analysis. The direction of the themes could also be looked at more in depth. For example, the statements "It was an interesting experience. I would not say it was a pleasant interaction." *p587* and "Quite negative, and it seemed Eliza just repeated what I said back but with a questionmark" *p623* both fall under broadly negative impressions, but suggest different levels of dissatisfaction. This nuance is not fully captured by the current thematic coding.

For the sub-questions, when aggregating the results across both prompt-based runs in the ***unguided prompt*** with all LLMs, a total of 23 out of 31 themes were identified, corresponding to a coverage rate of approximately 74%. This level of thematic coverage is notable, especially considering that none of the themes proposed by the LLMs were unrelated to the content of the responses. In other words, all identified themes were at least partially grounded in the source material, reinforcing the credibility of the model outputs. These findings suggest that LLMs may serve as a valuable supplementary tool in thematic analysis. Specifically, they can act as a secondary check to help researchers identify potentially overlooked themes and validate the completeness of manual coding efforts. This offers an additional layer of validation, wherein researchers can be reasonably

confident that the themes generated by the models are grounded in the source text. However, this was not the case for the ***guided prompt***, wherein the average kappa across models was low (Figure 6) with the overall mean average ($\kappa = 0.1438$) showing that LLMs are particularly bad at theme application. In summary, these findings show us that the manual thematic analysis remains the gold standard and only in the case of theme generation and not theme application LLMs can be used as a supplementary tool.

### 4.1 ASAQ

Our findings highlight that the ASAQ is a reliable, community-validated core instrument for capturing user experience with ASAs, as evidenced by the fact that 23 out of 31 (74%) of the themes identified through qualitative analysis could be directly mapped onto constructs from the ASAQ, supporting its practical usability and robustness. This suggests that the ASAQ covers a substantial portion of the range of experiences people refer to when discussing their interaction with ASAs. Furthermore, of the 23 themes 15 showed statistical significance, with a $p$-value $< 0.05$. Our thematic codes, when mapped to ASAQ constructs, also showed mostly positive correlations. This supports the validity of the constructs of the ASAQ, meaning that when the participants described something like trust, enjoyment or sociability in qualitative terms, it aligned with the quantitative measurement in the ASAQ. While no very strong correlations ($\rho > 0.9$) were observed, the general trend of the correlations being positive indicates consistent patterns of association with the ASAQ scores. This high degree of thematic alignment suggests that ASAQ effectively captures the core aspects users reflect on during ASA interactions, lending support to its potential as a standardized instrument. That said, this does also mean that 26% (8 out of 31) of the themes could not be directly mapped onto ASAQ constructs. Among these, *Agent's Limitation* stands out as the most noteworthy, having emerged in nearly 50 participant responses, suggesting that the usefulness of an agent for a limited purpose may represent a meaningful item not currently captured by the ASAQ. The remaining unmapped themes each had a notably lower response count, suggesting they may reflect more niche, context-dependent concerns.

### 4.2 Context within the Literature: ASAQ

To enhance the validity of our findings, we looked into the literature by integrating both qualitative thematic analysis and quantitative correlation analysis with the ASAQ framework and cross-verifying those results with other papers.

The general enjoyment observed in interactions with Artificial Social Agents (ASAs), as reflected in our findings seen in Figure 5, can also be derived from prior studies. For instance, Corrales-Paredes et. al. [26] found that participants generally enjoyed interacting with a robot agent and found it engaging. Also, consistent with our findings, delays and errors, as noticed by participant's engaging with *Eliza*, led to frustration. Their study also noticed that younger participants were less objective compared to older participants, which aligns with *Emotional Experience* being so high while simultaneously our dataset being underrepresented by older people. Additionally, their observation that some users perceived the robot as a mechanical tool rather than a sentient being corresponds with our *Human-like Behaviour* theme. However, their findings also suggest that a robot exhibiting more human-like behaviour tends to foster a more positive user perception. This contradicts our own findings, where the *Human-like Behaviour* didn't have as much of an influence of the positivity of other themes. This discrepancy may be attributed to our study not containing any humanoid robot agents.

While our dataset underrepresents older adults, Kim and Kim [27] conducted a focused study on a subset of this demographic. Their findings revealed that AI conversational agents provided both practical and emotional benefits to older adults living alone. Practically, the agents assisted with daily tasks, information retrieval, communication, and memory support. Emotionally, participants reported feeling happier, more secure and grateful. These outcomes align with patterns observed in our data as well, namely the *Agent's Helpfulness* and *Agent's Enjoyability* themes being perceived particularly high in positivity.

Lim et al. [28] report findings that differ markedly from ours. While the gender distribution across themes in our study remains relatively balanced (approximately 50/50, see Figure 4), their study observed a gender-based discrepancy, namely that female participants expressed a greater preference for VR-Embodied Conversational Agents (ECAs) than their male counterparts. Moreover, participants reported a stronger sense of presence when interacting with VR-embodied agents compared to text-only interfaces. These differences do unfortunately highlight an important gap in our current analysis, suggesting that the inclusion of Virtual Reality agents in future studies could reveal additional insights into user experience.

Similarly, in the context of Augmented Reality, Koleva et al. [29] emphasize the role of body language and non-verbal communication in shaping participants' perceptions of agents. This again points to a limitation in our agent set. Incorporating humanoid robots or embodied agents into future iterations of this study would provide a more holistic understanding of how interactions with Artificial Social Agents affect user perceptions and experiences.

### 4.3 Context within the Literature: The LLM Part

To further substantiate our results, we also looked into literature by comparing our findings with those reported in recent research on LLM-supported thematic analysis. Many recent research explores whether an LLM can identify the same experiences as manual analyses through mainly the human-in-the-loop frameworks [30, 31]. For example, Dai et al. [31] propose such a human-in-the-loop framework (called LLM-in-the-loop in the paper), in which the initial codes are generated by the LLM and subsequently refined through collaboration with a human coder. Their results show nearly perfect agreement between the human- and AI-Coder compared against the human coders, with Cohen's Kappa values of 0.87 and 0.81 for both respective

datasets. These results are higher than our findings, indicating that hybrid approaches involving close collaboration between humans and AI can likely outperform AI-only methods.

Similarly, Drápal et al. [16] also used an approach with and without an expert trying to improve the quality of the initial coding of the LLM. As they report, GPT-4 generated reasonable initial codes and was capable of refining them based on expert feedback. [16]. This observation resonates with our own findings, where LLMs produced sound initial codes and themes in the unguided prompt. Drápal et al. further state that 72.6% of the 785 predicted codes were initially deemed reasonable, closely aligning with our own results. Notably, that our models were able to identify around 74% of the themes previously found through manual analysis, supporting the assumption that LLMs can lead to effective coding outcomes. However, they also report an improvement to 88.8% following human refinement, which also underscores the potential of hybrid approaches.

Deiner et al. [32] take a broader view, examining whether LLMs can replicate or approximate human-conducted thematic analysis in health-related social media data. They found that LLMs identified several themes similar to those generated by human analysts, with low hallucination rates. However, variability was observed both between different LLMs and between test runs of an individual LLM. Although the LLM-generated themes did not consistently match the human-generated themes, subject matter experts still considered them reasonable and relevant. Similarly, in our study, we observed low hallucination rates and found that themes generated by the LLMs were reasonable and relevant. Importantly, although LLMs were found to produce relevant and reasonable themes, they did not consistently match the depth or specificity of those generated by human analysts. This is also indicated in our guided prompt, which resulted in low average $\kappa$ across themes. The authors conclude that while LLMs show promise for large-scale, real-time thematic extraction, particularly in public health applications, they are not yet capable of fully replicating human-level analysis. This conclusion would be one that is also shared and supported by our own findings.

### 4.4 Limitations

There were some limitations as to the dataset, as seen in Figure 4, namely that the dataset is not that indicative of older age-groups. Furthermore, there was an under representation of people with no formal education in the datasets. As a result, the conclusions drawn from the study may be less applicable to demographic groups that were not well represented. So, to ensure that this limitation is somewhat taken care of, we cross-validated results from papers specifically targeting these groups. Also, as stated in the responsible research section, a single Inter-Coder Agreement is not enough to take the conclusion as-is and more should be done to reduce the bias to an absolute minimum. Another limitation is that participants were compensated for contributing to the dataset, which may have influenced the authenticity and depth of their responses. Since payment was not contingent on the quality or effort of their input, some individuals may have provided minimal or insincere responses simply to complete the task faster, potentially affecting the overall quality and reliability of the data. With regards to the agents, they were not fully representative of all possible agents. Notably, a humanoid agent was missing. With regards to the interaction, the environments played a role in the interaction and not all environments were presented to the users, namely augmented and virtual reality. Finally, Demographic distribution should be noted as a limitation, especially the underrepresentation of some regions, which might influence the generalizability of our results. Future work could explore these themes with more diverse samples or focus on improving aspects that receive critical feedback.

### 4.5 Future Work

Thematic analysis is a qualitative method used to identify, analyze, and report themes (read *patterns*) within data. When we think about recognizing patterns, the first thing that often comes to mind is machine learning. By systematically identifying themes through thematic analysis, these patterns could potentially be used to train machine learning algorithms for automated recognition. Exploring this intersection could be a promising direction for future research. To reflect further on this point, rather than LLMs, other potential automatic approaches can be used to perform the thematic analysis. One such approach is active learning, with *small-text* [33, 34] as a candidate library to use for further research. The results then can be compared to the already obtained results from this study to further enforce or perhaps disprove them. To perhaps state the obvious, more people could do the thematic analysis, more data could be gathered to do the analysis on and stronger LLMs can be used for the analysis. Another interesting avenue could be looking into conducting the analysis within augmented or virtual reality. Looking into whether this takes care of certain concerns or even shifts the directions -positive or negative- of certain themes. Finally, Semantic Analysis can be conducted on the dataset, with libraries such as TextBlob [35], to either substantiate or refute the general sentiment derived from our thematic analysis.

# 5 Responsible Research

## 5.1 Ethical Considerations

The study is based on a pre-existing dataset of user interactions with Artificial Social Agents. The dataset was originally collected under appropriate ethical guidelines. No personal identifiable information was present in the dataset. The only information present and used were anonimized descriptors of each participant.

## 5.2 Transparency and Reproducibility

We aimed to make this study fully transparent and reproducible. The dataset, cleaned up, with all the keywords, codes and themes and the entire process is present in an ODS file on the Github repository [25]. In the same repository the Inter-Coder Agreements and the Python code has been put, such that future researchers can view and reproduce the process or results as much as possible. Mistakes made during the process can also be noted and corrected. The repository is free to use for everyone with a permissive MIT license.

## 5.3 Large Language Models

The Large Language models used for the data were all local, thus the dataset itself was not used as training for the LLM models. This ensures data safety and data privacy. For reproducibility, a table was created with all LLMs and their versions used (see Table 7). Furthermore, using local LLMs ensures that the models remain consistent and transparent, without the filtering, updates, or potential manipulation associated with online models like ChatGPT, which often operate as opaque 'black boxes'.

Table 7: Large Language Models (LLMs) and their versions used in the study. In all models the quant was **Q4_K_M**.

| Publisher | Params | Model |
|---|---|---|
| AliBaBa Qwen | 32B | qwen3-32b |
| DeepSeek R1 Distill | 32B | deepseek-r1-distill-qwen-32b |
| Microsoft Phi | 15B | phi-4 |
| Mistral NeMo | 12B | mistral-nemo-instruct-2407 |
| Google Gemma | 12B | gemma-3-12b |
| Meta Llama | 8B | llama-3.1-8b-instruct |

## 5.4 Social Impacts, Risks and Biases

This study should not be taken as is, and the conclusions should be considered within its context and the researchers' biases. It was made by Computer Science students as their thesis project. This ensures that there is probably a bias towards code-ability of the data and results. This could ensure an over-reliance on objectivity or quantitative methods. This also means, generally speaking, a limited training in qualitative methods, unfortunately. Also, the data and users' feedback is interpreted through the lens of a software developer. For example, making datasets more code-able for computer-analysis is one of such biases derived from it. The usage of jargon in a "techy" way is another. The use of such language could ensure a technical-centric viewpoint. Moreover, the obvious biases in gender, age-group, cultural/ethnic background and religious beliefs would also play a role.

Furthermore, as discussed before, the collected dataset contains several biases. For example, the agents used were not fully indicative of all possible agents, i.e. a humanoid robot was not present. It was also heavily skewed towards relatively younger people, with some form of education other than a PhD. As a final remark, the thematic analysis was done by a single researcher and a sample of it was given towards a single researcher to do the bias-check. While this step adds a level of validation, it is not sufficient to fully eliminate potential subjectivity. The transparency of the entire process, does ensure that others can build upon it and more analyses can be made towards it, reducing this subjectivity even further.

## 5.5 Fair and Open Data

All libraries used use permissive open-source licenses, namely BSD, MIT and the Python Software Foundation License. The code and data itself is present at a public Github repository [25]. All prompts used for the LLM analysis can be found in Appendix A or at the previously mentioned Github repository.

## 6 Conclusion

This study set out to explore how individuals experience interactions with Artificial Social Agents (ASAs), which was done by identifying recurring themes through manual thematic analysis, and assess the viability of Large Language Models (LLMs) as tools for automating this process. Through an iterative manual coding process, 31 distinct themes were identified, offering insight into what users value, expect, and critique in their interactions with ASAs. This study found that user interactions with Artificial Social Agents (ASAs) are generally perceived as positive, interesting and valuable. Notably, themes such as *Agent's Helpfulness*, *Attitude*, and *Human-like Behaviour* stood out as central to the user experience as they were the most found themes. These findings suggest that people seek not only practicality but also emotional resonance in their interactions with ASAs. Among these, The *Human-like Behaviour* theme revealed a tension: while some users appreciated anthropomorphic traits and found them comforting or engaging, others were put off by them. ASAs were primarily seen as practical tools that enhanced users' daily lives through improved productivity, autonomy, and ease of use. Agents with friendly tones and warm sociability were received more positively, while cold or robotic responses detracted from the experience. Users responded positively to agents that felt helpful, enjoyable, and sociable, highlighting a growing expectation for these systems to go beyond basic task execution and engage in more human-centered, meaningful exchanges. Most users saw ASAs primarily as tools, however a few saw them as companions.

In addition, themes like *Agent's Cognition*, *Coherence*, and *Intentionality* show that users assess intelligence of the agent, from which a mixed sentiment was derived: people praised insightful responses but quickly noticed incoherence, contradictions or awkward behaviour which led to dissatisfaction. Furthermore, themes that focused on how the agent influenced the user's daily routine, mood, and willingness to reuse it, (e.g. *Ease of Life*, *Productivity*, *Autonomy*, *User Acceptance*) had positive experiences which often led users to express a desire to continue using the agent. Also, correlations between the themes and the 90 constructs of the ASAQ showed positive alignment in the majority of themes, particularly for themes like *Agent's Enjoyability*, *Sociability*, *Reliability*, and *Cognition*. This reinforces the credibility of all the themes and strengthens our findings of our thematic analysis. Overall, most users expressed generally positive experiences across agents, except for older agents like *Eliza*, which were often viewed as outdated, frustrating or awkward.

In addition to the manual analysis, this study evaluated the capabilities of various locally hosted Large Language Models (LLMs) in conducting thematic analysis. We demonstrated that while locally hosted LLMs were capable of capturing general thematic structures through unguided prompts—achieving a 74% overlap with manually identified themes, they performed poorly in the guided, response-level thematic annotation task. The low Inter-Coder Agreement (as measured by Cohen's Kappa) revealed that current LLMs lack the consistency and nuance required for fine-grained qualitative analysis, particularly when applying predefined coding frameworks. The high overlap does suggest, however, that LLMs can serve as valuable auxiliary tools in qualitative research, particularly in the early stages of theme discovery or as a secondary check for human-led analysis.

## 7 Acknowledgment

During the writing process, we used OpenAI's ChatGPT as described in Appendix F. The actual content, analysis and interpretations presented in this paper are entirely our own. Furthermore, we would like to express our gratitude towards our supervisor, Willem-Paul Brinkman, for their guidance, feedback and support throughout. We would also like to extend our gratitude to our peer researcher, Antonio Lupu, conducting the peer analyses used throughout our paper. Finally, we acknowledge all fellow research group members who generously contributed their time and perspectives to this study: Keshav Nair, Antonio Lupu, Jason Miao and Andreea Ţebrean.

## References


[1] Leon Ciechanowski et al. “In the shades of the uncanny valley: An experimental study of human–chatbot interaction”. In: *Future Generation Computer Systems* 92 (2019), pp. 539–548. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2018.01.055. URL: https://www.sciencedirect.com/science/article/pii/S0167739X17312268.

[2] Janusz Kacprzyk and Slawomir Zadrozny. “Computing with words is an implementable paradigm: Fuzzy queries, linguistic data summaries, and natural-language generation”. In: *IEEE Transactions on Fuzzy Systems* 18.3 (2010), pp. 461–472.

[3] Siska Fitrianie et al. “What are We Measuring Anyway? - A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences”. In: *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. IVA ’19. Paris, France: Association for Computing Machinery, 2019, pp. 159–161. ISBN: 9781450366724. DOI: 10.1145/3308532.3329421. URL: https://doi.org/10.1145/3308532.3329421.

[4] Virginia Braun and Victoria Clarke and. “Using thematic analysis in psychology”. In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa. URL: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

[5] Saul McLeod. *Thematic Analysis: A Step by Step Guide*. June 2024. DOI: 10.13140/RG.2.2.13084.71048.

[6] Maximo R Prescott et al. “Comparing the Efficacy and Efficiency of Human and Generative AI: Qualitative Thematic Analyses”. In: *JMIR AI* 3 (Aug. 2024), e54482. ISSN: 2817-1705. DOI: 10.2196/54482. URL: https://doi.org/10.2196/54482.

[7] Walter S. Mathis et al. “Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: how does it compare to traditional methods?” English. In: *Computer Methods and Programs in Biomedicine* 255 (Oct. 2024). © 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies. Data availability statement: Not present. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2024.108356.

[8] Takeshi Kondo et al. “A mixed-methods study comparing human-led and ChatGPT-driven qualitative analysis in medical education research”. In: *Nagoya Journal of Medical Science* 86.4 (2024), pp. 620–644. DOI: 10.18999/nagjms.86.4.620. URL: https://doi.org/10.18999/nagjms.86.4.620.

[9] Siska Fitrianie et al. “The artificial-social-agent questionnaire: Establishing the long and short questionnaire versions”. English. In: *IVA 2022 - Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. IVA 2022 - Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents. 22nd ACM International Conference on Intelligent Virtual Agents, IVA 2022 ; Conference date: 06-09-2022 Through 09-09-2022. United States: ACM, 2022. DOI: 10.1145/3514197.3549612.

[10] Siska Fitrianie et al. “The Artificial Social Agent Questionnaire (ASAQ) - Development and evaluation of a validated instrument for capturing human interaction experiences with artificial social agents”. In: *International Journal of Human-Computer Studies* 199 (2025), p. 103482. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2025.103482. URL: https://www.sciencedirect.com/science/article/pii/S1071581925000394.

[11] Willem Paul Brinkman. *The Artificial Social Agent Questionnaire (ASAQ) tutorial*. 2025. URL: https://www.youtube.com/watch?v=n5qypQbrBPk.

[12] Rensis Likert. “A Technique for the Measurement of Attitudes”. In: *Archives of Psychology* 22.140 (1932), pp. 1–55.

[13] Virginia Braun and Victoria Clarke. “Supporting best practice in reflexive thematic analysis reporting in Palliative Medicine: A review of published research and introduction to the Reflexive Thematic Analysis Reporting Guidelines (RTARG)”. In: *Palliative Medicine* 38.6 (2024), pp. 608–616. DOI: 10.1177/02692163241234800. eprint: https://doi.org/10.1177/02692163241234800. URL: https://doi.org/10.1177/02692163241234800.

[14] Virginia Braun and Victoria Clarke. *Thematic analysis: Choosing a suitable approach*. Accessed: 2025-06-02. 2021. URL: https://the-sra.org.uk/SRA/SRA/Blog/ThematicanalysisChoosingasuitableapproach.aspx.

[15] David Byrne. “A Worked Example of Braun and Clarke’s Approach to Reflexive Thematic Analysis”. In: *Quality & Quantity* 56.3 (June 2022), pp. 1391–1412. ISSN: 1573-7845. DOI: 10.1007/s11135-021-01182-y.

[16] Jakub Drápal, Hannes Westermann, and Jaromir Savelka. *Using Large Language Models to Support Thematic Analysis in Empirical Legal Studies*. 2023. arXiv: 2310.18729 [cs.AI]. URL: https://arxiv.org/abs/2310.18729.

[17] Jacob Cohen. “A coefficient of agreement for nominal scales”. In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.

[18] J. Richard Landis and Gary G. Koch. “The Measurement of Observer Agreement for Categorical Data”. In: *Biometrics* 33.1 (1977), pp. 159–174. DOI: 10.2307/2529310. URL: https://doi.org/10.2307/2529310.

[19] G. G. Simpson. “Mammals and the Nature of Continents”. In: *American Journal of Science* 241.1 (1943), pp. 1–31.

[20] Willem Paul Brinkman. *The Artificial Social Agent Questionnaire (ASAQ) website*. 2025. URL: https://ii.tudelft.nl/evalquest/web/node/1.

[21] Haldun Akoglu. "User's guide to correlation coefficients". In: *Turkish Journal of Emergency Medicine* 18.3 (2018), pp. 91–93. ISSN: 2452-2473. DOI: https://doi.org/10.1016/j.tjem.2018.08.001. URL: https://www.sciencedirect.com/science/article/pii/S2452247318302164.

[22] Christine Dancey and John Reidy. *Statistics Without Maths for Psychology (7th edition)*. May 2017. ISBN: 9781292128856.

[23] Joseph Weizenbaum. "ELIZA—A Computer Program For the Study of Natural Language Communication Between Man and Machine". In: *Communications of the ACM* 9.1 (1966), pp. 36–45. URL: https://doi.org/10.1145/365153.365168.

[24] Patrick Schober, Christa Boer, and Lothar A. Schwarte. "Correlation Coefficients: Appropriate Use and Interpretation". In: *Anesthesia & Analgesia* 126.5 (May 2018), pp. 1763–1768. DOI: 10.1213/ANE.0000000000002864.

[25] Celal Karakoç. *Interaction with Artificial Social Agents - A thematic analysis of people's experiences*. [Source Code and Data]. 2025. URL: https://github.com/ckarakoc/bep-asa.

[26] Ana Corrales-Paredes et al. "User Experience Design for Social Robots: A Case Study in Integrating Embodiment". In: *Sensors* 23.11 (2023). ISSN: 1424-8220. URL: https://www.mdpi.com/1424-8220/23/11/5274.

[27] Kyung Mee Kim and Sook Hyun Kim. "Experience of the Use of AI Conversational Agents Among Low-Income Older Adults Living Alone". In: *SAGE Open* 14.4 (2024), p. 21582440241301022. DOI: 10.1177/21582440241301022. URL: https://doi.org/10.1177/21582440241301022.

[28] Sue Lim, Ralf Schmälzle, and Gary Bente. *Artificial social influence via human-embodied AI agent interaction in immersive virtual reality (VR): Effects of similarity-matching during health conversations*. 2024. arXiv: 2406.05486 [cs.HC]. URL: https://arxiv.org/abs/2406.05486.

[29] Katerina Koleva et al. *Influence of Personality and Communication Behavior of a Conversational Agent on User Experience and Social Presence in Augmented Reality*. 2024. arXiv: 2403.09883 [cs.HC]. URL: https://arxiv.org/abs/2403.09883.

[30] Huimin Xu et al. *TAMA: A Human-AI Collaborative Thematic Analysis Framework Using Multi-Agent LLMs for Clinical Interviews*. 2025. arXiv: 2503.20666 [cs.HC]. URL: https://arxiv.org/abs/2503.20666.

[31] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. "LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9993–10001. DOI: 10.18653/v1/2023.findings-emnlp.669. URL: https://aclanthology.org/2023.findings-emnlp.669/.

[32] Michael S Deiner et al. "Large Language Models Can Enable Inductive Thematic Analysis of a Social Media Corpus in a Single Prompt: Human Validation Study". In: *JMIR Infodemiology* 4 (Aug. 2024), e59641. ISSN: 2564-1891. DOI: 10.2196/59641. URL: https://doi.org/10.2196/59641.

[33] small-text. *small-text: Active Learning for Text Classification in Python*. 2025. URL: https://github.com/webis-de/small-text.

[34] Christopher Schröder et al. "Small-Text: Active Learning for Text Classification in Python". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 84–95. URL: https://aclanthology.org/2023.eacl-demo.11.

[35] Steven Loria. *TextBlob: Simplified Text Processing*. 2018. URL: https://textblob.readthedocs.io.

# Appendix

## A LLM Prompts

### Unguided Prompt: Given the text, give me the Themes.

```
You are an expert qualitative researcher conducting a thematic analysis to answer the
    Research Question: "How do people experience their interaction with Artificial Social
    Agents?"

Please perform a full thematic analysis of the following text, following these three steps
    carefully:

Step 1: Familiarization
- Read the entire text thoroughly.
- Provide a concise summary of the overall topics, key impressions, and recurring ideas found
     in the data (3-5 sentences).

Step 2: Coding
- Identify the most relevant codes (labels) in the text, each named in ≤3 words.
- Present as a table with columns: | Code | Description |
- The Description should clearly explain what the code represents.

Step 3: Grouping
- Organize the identified codes into coherent themes or groups based on conceptual similarity
     or relation.
- Present as a table with columns: | Group # | Theme | Description |
- The Description should briefly explain the theme and how it relates to the research
    question.

Here is the full text to analyze:
---{Input=}
Response_0
...
Response_n
---

Input: All user-responses seperated by a line-break.
Output: A Summary. A Table of Codes. A Table of Themes.
```

### Guided Prompt: Given the themes, can the LLM find them in the sample.

```
You are an expert qualitative researcher conducting a thematic analysis, trying to answer the
     research question: "How do people experience their interaction with Artificial Social
    Agents?".
Analyze the following themes and their definition:
---
[Theme_0]: [Definition_0]
...
[Theme_n]: [Definition_n]
---

Analyze the following user responses and identify all applicable themes based on the provided
     definitions. Only assign a theme if there is clear and explicit support for it in the
    response - do not infer or assume. Give me a comma-separated list of all themes that you
    find in the following user-responses to an online questionnaire:
---{Input=}
User_0: Response_0
...
User_n: Response_n
---
Present as a table with columns: | Response # | Themes |
Do not include any explanations or additional text outside of the table itself.

Input: A sample of 100 user-responses seperated by a line-break.
Output: A Table of Themes found per Response.
```

## B Thematic Analysis

Table 8: Final themes and example quotes

| Theme | Quote |
|---|---|
| **Agent's Cognition** | "I find Alexa really useful, although sometimes Alexa does get things wrong or doesn't hear/understand what Im asking for" *p56* |
| **Agent's Coherence** | "Its a great AI but it doesn't always provide the answers I want, so I would have to opt for another AI." *p89* |
| **Agent's Creativeness** | "I use it mostly for work, when I'm having difficulty with a task I ask for advice or workarounds. Also, creatively it can provide me with different inspirations." *p133* |
| **Agent's Efficiency** | "Using ChatGPT makes my work (life) so much easier. It cuts time spent doing research or editing a document significantly." *p81* |
| **Agent's Emotional Presence** | "The experience was awkward, Eliza responded by mimicking and being rude. I disliked the conversation and it was not helpful or interesting." *p663* |
| **Agent's Enjoyability** | "I enjoyed it" *p92* |
| **Agent's Helpfulness** | "It was a wonderful experience because Siri was helpful especially when I was busy and late I would ask Siri to check the weather for me instead of searching om my phone." *p10* |
| **Agent's Intentionality** | "Eliza created replies based on my sentences but didn't seem to parse the meaning of what I had said, rather she inverted by statements to create questions that were absent of context" *p362* |
| **Agent's Interestingness** | "A very interesting experience. I felt important to her and I felt that I was not alone." *p14* |
| **Agent's Intuitiveness** | "i have had a pleasant experience with Alexa, its easy to use and intuituve" *p461* |
| **Agent's Limitation** | "It doesn't help me much with university work but It helps me with hobbies like textual role-playing. " *p30* |
| **Agent's Personality** | "I believe it was a bit bizarre as both i and Eliza were not saying things that made much sense, seemed like it would just make very generic answers as a person who's barely listening" *p455* |
| **Agent's Quickness** | "I use regularly for many reasons and it is very reliable and fast." *p107* |
| **Agent's Reliability** | "alexa is knowledable and I can ask simple staff but can not rely 100% on her, also sometimes responds in funny manner but still its AI" *p116* |
| **Agent's Sociability** | "copilot was a really fun experience , it was really knowledgeable and it helped me out a lot, it also seemed like i had company with me" *p335* |
| **Agent's Usability** | "I use Alexa to help me do easy tasks like searching definitions, do simple tasks like playing music and I think it's easy to use" *p4* |
| **Attitude** | "I used It to ask some silly questions like how to pick up older women etc, It was funny and informative" *p13* |
| **Ease of Life** | "It has been a very pleasant experience and has made my professional life way easier." *p196* |
| **Emotional Experience** | "my experience was average, nothing extremely interesting i was not amazed" *p48* |
| **Human-like Behaviour** | "it doesn't really have anything special, i didn't feel a strong connection it was just like talking with a robot" *p37* |
| **Limitations** | "I asked for some easily verifiable information and it gave correct and seemingly well though out answers. However when I asked for the source of this information, it produced publications which don't even exist.<br>It also seemed to be no nearly useless when faced with some math problems." *p24* |
| **Performance** | "It was fine. I was able to find some functions I wasn't aware that are possible I was humming a song and it gave me a list of a songs that I might be thinking of" *p44* |
| **Potential** | "Alexa is interactive but lacks emotion but sometimes it feels like as if we are talking with a real person. Alexa will be beneficial for the future." *p471* |
| **Productivity** | "i think its a good idea by using Alexa on a daily basis as it helps increasing my productivity and of course i am going to continue using Alexa in the future." *p72* |

*Continued on next page*

| Theme | Quote |
|---|---|
| **User Acceptance** | "I felt really strange because I have never used anything like this, but after that I still don't think that I need it. My opinion about this hasn't changed" *p300* |
| **User-Agent Alliance** | "I've mainly used ChatGPT to simply test out the technology. I do find it really interesting that it can generate a significant amount of text that is both coherent and somewhat creative in a short amount of time. While I haven't had the need to use the technology in a professional setting, I will be keeping an eye on it to see what applications it could have in my field." *p12* |
| **User-Agent Interplay** | "I use it mostly for work, when I'm having difficulty with a task I ask for advice or workarounds. Also, creatively it can provide me with different inspirations." *p133* |
| **User's Autonomy** | "I would describe it as a personal assistant to help me keep up with the house work, so I can work on other task as it helps me." *p218* |
| **User's Emotional Presence** | "I did not enjoy using Eliza at all. I found it very hard to read her tone and she came off very abrasive and accusatory. I felt worse after talking to Eliza than I did before I started" *p664* |
| **User's Engagement** | "I'm not to keen on our interaction. I value its practicality but it's not for me." *p46* |
| **User's Trust** | "i felt that I could trust the robot to assist bringing up issues people in the family might have but are not comfortable to talk about.it was a very positive experience" *p20* |

## C Peer Thematic Analysis

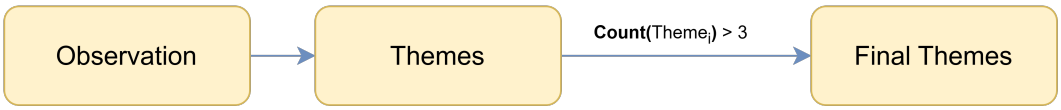


Figure 8: Thematic Analysis approach used by the peer.

Table 9: Main themes of the Peer and Definitions

| Theme (Peer) | Definition |
|---|---|
| **Accessibility** | How easy it is to use for people with different abilities or needs |
| **Accuracy** | How often it gives correct or reliable answers |
| **Convenience** | How much it makes your life easier or saves you time |
| **Creativity** | How well it comes up with new or interesting ideas |
| **Efficiency** | How quickly and effectively it helps get things done |
| **Emotional Connection** | Whether the user feels any personal connection with it or not |
| **Engagement** | How much users are engaged with it |
| **Enjoyability** | How much users like or enjoy using it |
| **Entertainment** | How fun it is to interact with |
| **Helpfulness** | How much it helps solve problems or answer questions |
| **Human-Like Behavior** | How much it feels like talking to or working with a person |
| **Interestingness** | How much users find it interesting |
| **Limitations** | User thoughts on things it cannot do well or problems/limitations noticed |
| **Potential** | Thoughts on how useful the tool could be in the future, how it could improve, or excitement about future use |
| **Productivity** | How much it helps the user improve their workflow |
| **Trust** | How much users feel they can rely on it to give good answers or do the right thing |
| **Usability** | How easy it is to use and interact with |

Table 10: Final mapping of themes after the Inter-Coder Agreement on the manual thematic analysis.

| **Theme (Coder 1)** | **Theme (Coder 2)** |
|---|---|
| Agent's Cognition | – |
| Agent's Coherence | Accuracy |
| Agent's Creativeness | Creativity |
| Agent's Efficiency | Efficiency |
| Agent's Emotional Presence | – |
| Agent's Enjoyability | Enjoyability |
| Agent's Helpfulness | Helpfulness |
| Agent's Intentionality | – |
| Agent's Interestingness | Interestingness |
| Agent's Intuitiveness | – |
| Agent's Limitation | – |
| Agent's Personality | – |
| Agent's Quickness | – |
| Agent's Reliability | – |
| Agent's Sociability | – |
| Agent's Usability | Usability, Accessibility, Convenience |
| Attitude | Entertainment |
| Ease of Life | – |
| Emotional Experience | Emotional Connection |
| Human-like Behaviour | Human-like Behavior |
| Performance | – |
| Potential | Potential |
| Productivity | Productivity |
| User Acceptance | – |
| User's Autonomy | – |
| User's Emotional Presence | – |
| User's Engagement | Engagement |
| User's Trust | Trust |
| User-Agent Alliance | – |
| User-Agent Interplay | – |
| – | Limitations |

Table 11: Calculations of Cohen's Kappa on mapped themes.

| **Theme** | $\kappa$ | **Interpretation** $\kappa$ | **a** | **b** | **c** | **d** |
|---|---|---|---|---|---|---|
| Agent's Coherence | 0.83 | Almost perfect agreement | 8 | 2 | 1 | 89 |
| Agent's Creativeness | 0.92 | Almost perfect agreement | 6 | 0 | 1 | 93 |
| Agent's Efficiency | 0.93 | Almost perfect agreement | 7 | 0 | 1 | 92 |
| Agent's Enjoyability | 0.71 | Substantial agreement | 7 | 5 | 0 | 88 |
| Agent's Helpfulness | 0.79 | Substantial agreement | 58 | 5 | 5 | 32 |
| Agent's Interestingness | 0.28 | Fair agreement | 3 | 13 | 0 | 84 |
| Agent's Usability | 0.8 | Almost perfect agreement | 24 | 6 | 2 | 68 |
| Attitude | 0.2 | Fair agreement | 4 | 21 | 1 | 74 |
| Emotional Experience | 0.33 | Fair agreement | 15 | 32 | 0 | 53 |
| Human-like Behaviour | 0.5 | Moderate agreement | 9 | 14 | 0 | 77 |
| Potential | 0.65 | Substantial agreement | 3 | 3 | 0 | 94 |
| Productivity | 0.74 | Substantial agreement | 3 | 2 | 0 | 95 |
| User's Engagement | 0.63 | Substantial agreement | 7 | 7 | 0 | 86 |
| User's Trust | 0.71 | Substantial agreement | 4 | 3 | 0 | 93 |

# D Quantitative Analysis

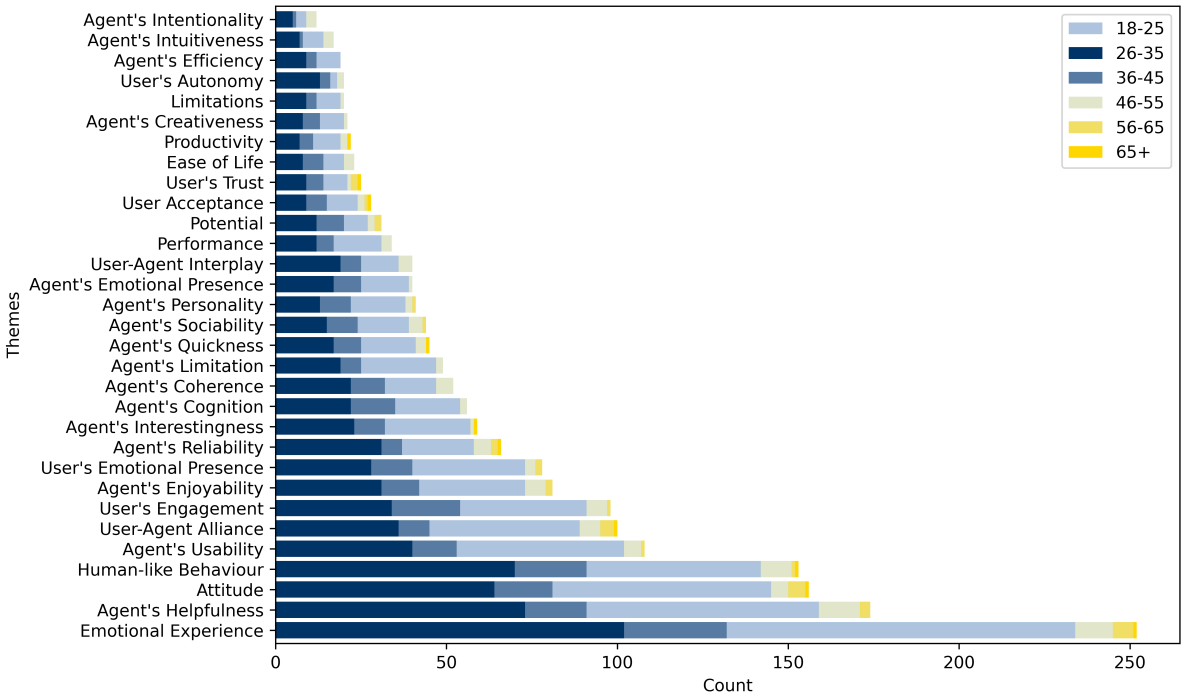


Figure 9: Themes by Age Group

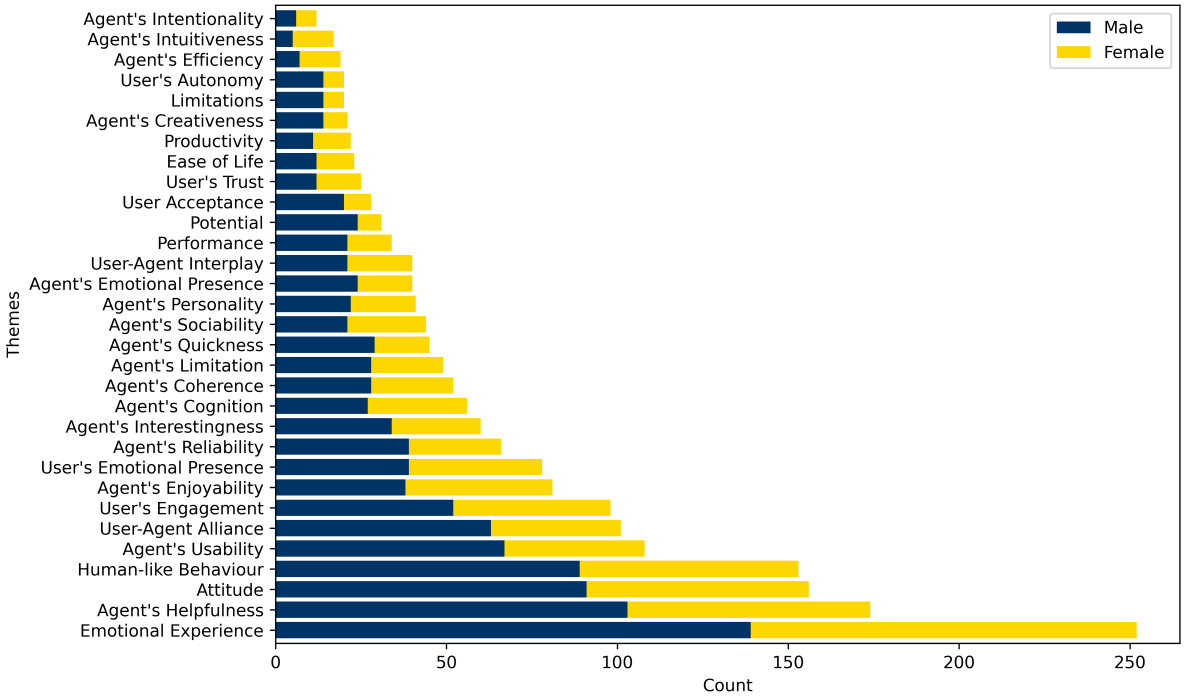


Figure 10: Themes by Gender

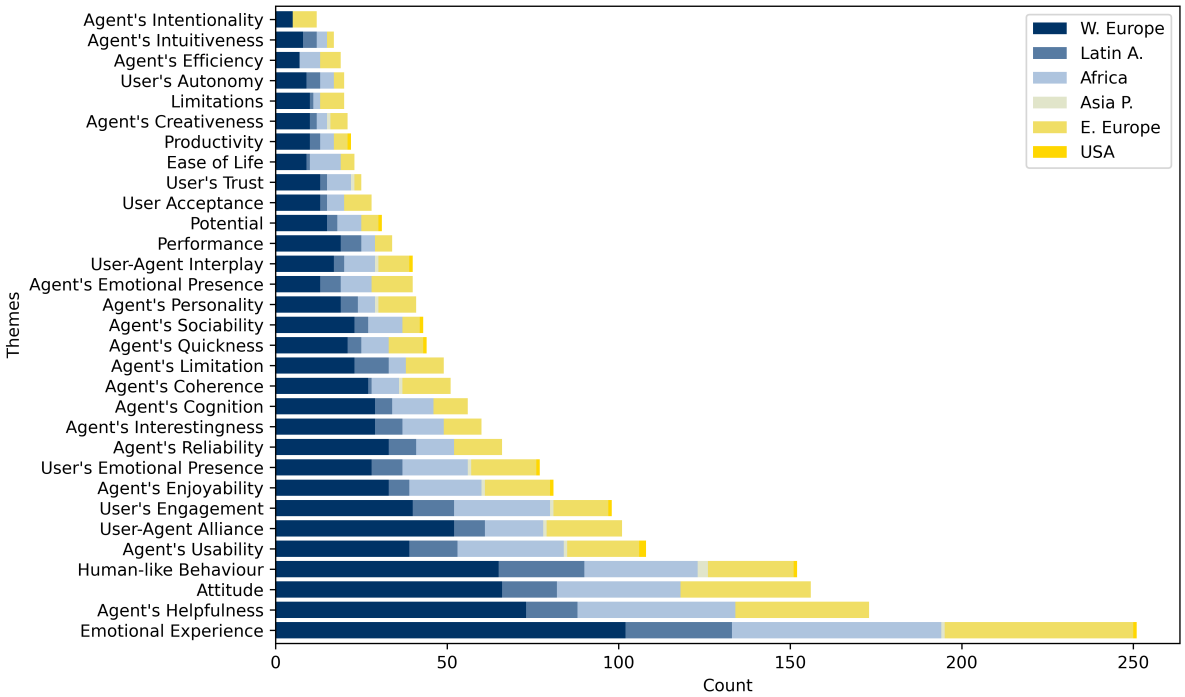


Figure 11: Themes by Region

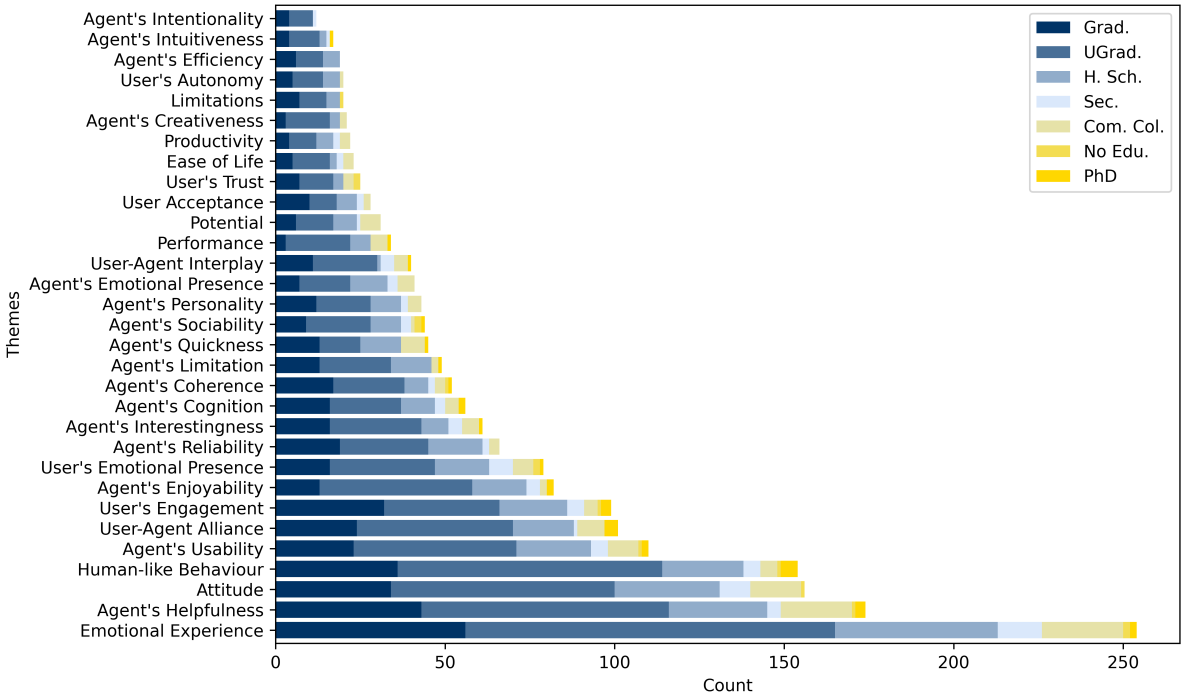


Figure 12: Themes by Education

Table 12: Count of themes found in the dataset

| **Themes** | **Count** |
|---|---|
| Emotional Experience | 254 |
| Agent's Helpfulness | 174 |
| Attitude | 157 |
| Human-like Behaviour | 154 |
| Agent's Usability | 110 |
| User-Agent Alliance | 101 |
| User's Engagement | 99 |
| Agent's Enjoyability | 82 |
| User's Emotional Presence | 79 |
| Agent's Reliability | 66 |
| Agent's Interestingness | 61 |
| Agent's Cognition | 56 |
| Agent's Coherence | 52 |
| Agent's Limitation | 49 |
| Agent's Quickness | 45 |
| Agent's Sociability | 44 |
| Agent's Personality | 43 |
| Agent's Emotional Presence | 41 |
| User-Agent Interplay | 40 |
| Performance | 34 |
| Potential | 31 |
| User Acceptance | 28 |
| User's Trust | 25 |
| Ease of Life | 23 |
| Productivity | 22 |
| Agent's Creativeness | 21 |
| Limitations | 20 |
| User's Autonomy | 20 |
| Agent's Efficiency | 19 |
| Agent's Intuitiveness | 17 |
| Agent's Intentionality | 12 |

## E LLM Thematic Analysis

Table 13: Model Comparison by Average $\kappa$

| Model | Avg. $\kappa$ | Interpretation |
|---|---|---|
| Llama | 0.0042 | Slight agreement |
| Qwen | 0.1409 | Slight agreement |
| DeepSeek | 0.2728 | Fair agreement |
| Phi | 0.1446 | Slight agreement |
| Gemma | 0.2104 | Fair agreement |
| NeMo | 0.0897 | Slight agreement |

## F LLM Usage within Paper

For the paper itself ChatGPT has been used to improve sentence structure of already written parts. As an example:

| **Initial** | | **Suggestion** | | **Result** |
|---|---|---|---|---|
| One such challenge is its recursive nature, in which researchers move back and forth between the phases of thematic analysis. | $\Rightarrow$ | One such difficulty lies in its recursive nature, which requires researchers to iteratively move between different phases of the analysis. | $\Rightarrow$ | One such challenge is its recursive nature, which ensures that researchers iteratively move between different phases of the analysis. |

ChatGPT was also used for suggesting transitional words or synonyms, improving the wording and flow of the sentences. Furthermore, AI has been used in the coding process, mainly the AI already present in the JetBrains IDEs and ChatGPT.

## G Formulas

### Spearman's Correlation

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$d_i$ = difference between the ranks of each observation

$n$ = number of paired observations

### Cohen's Kappa

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \quad P_o = \frac{a + d}{N}, \quad P_e = \left( \frac{a + b}{N} \cdot \frac{a + c}{N} \right) + \left( \frac{c + d}{N} \cdot \frac{b + d}{N} \right)$$

$N$ = Total responses = $a + b + c + d$

$P_o$ = Observed agreement

$P_e$ = Expected agreement by chance

$a$ = Agreement: theme present in both

$b$ = Disagreement: theme present in only one (us, but not the peer)

$c$ = Disagreement: theme present in only one (the peer, but not us)

$d$ = Agreement: theme present in neither