Interaction with Artificial Social Agents a thematic analysis of people's experiences

Research Questions

[RQ1] - How do people experience their interaction with Artificial Social Agents?

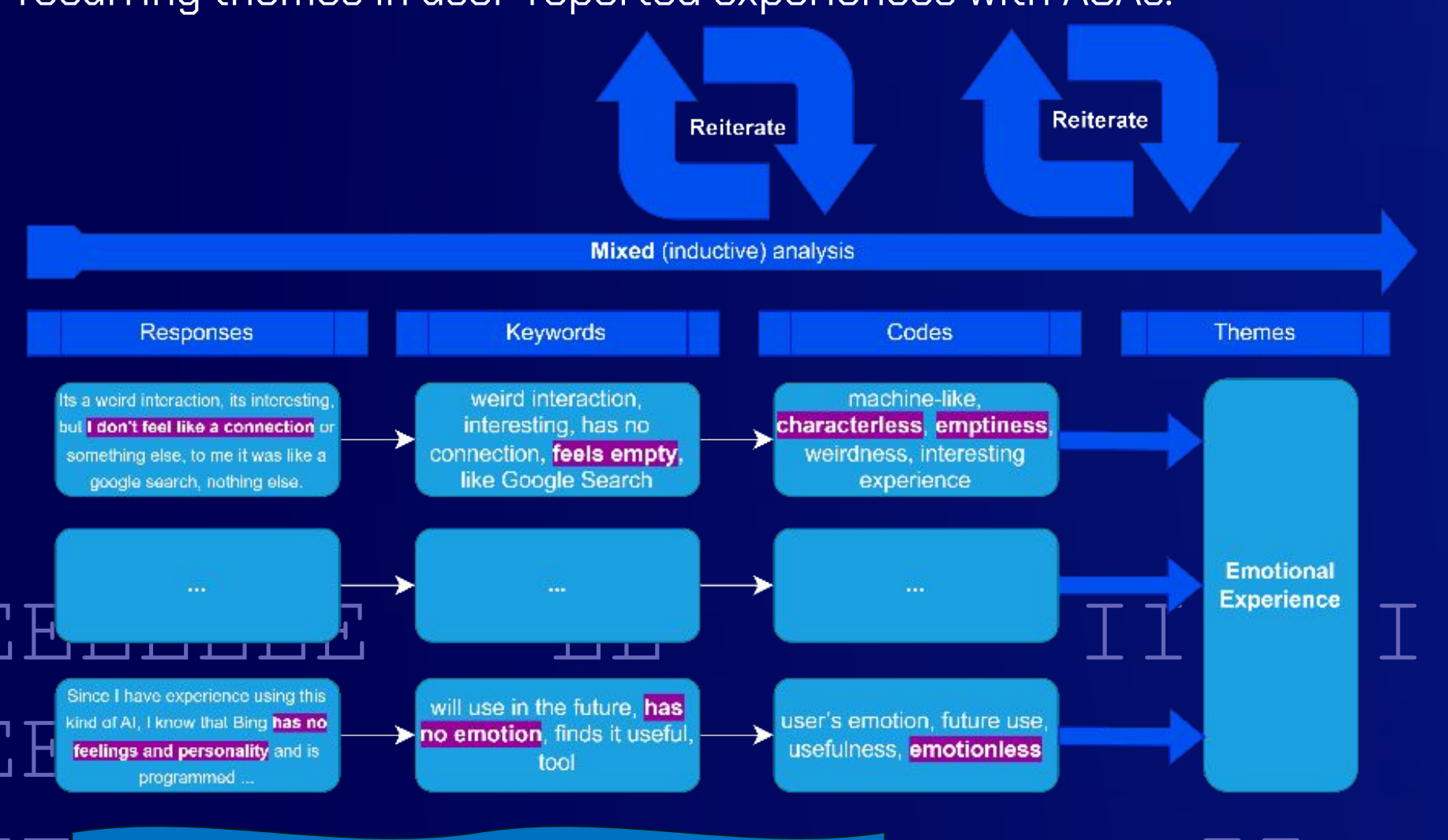
[SQ1] - Can a (locally hosted) Large Language Model (LLM) identify these experiences?

[SQ2] - How do manual and LLM-based thematic analysis compare with each other?

Research

The use of Artificial Social Agents is rapidly expanding across society. As these agents become more integrated into our interactions, understanding the user perception and experience of them becomes increasingly necessary to ensure their design aligns with user needs, promotes trust, and supports meaningful engagement.

- The aim of this study is to investigate how users experience interactions with Artificial Social Agents (ASAs), focusing on using **thematic analysis** to identify recurring themes in user-reported experiences with ASAs.



Reiterate
Reiterate
Mixed (inductive) analysis
Responses
Keywords
Codes
Themes
Its a weird interaction, its interesting, but I don't feel like a connection or something else, to me it was like a google search, nothing else.
weird interaction, interesting, has no connection, feels empty, like Google Search
machine-like, characterless, emptiness, weirdness, interesting experience
...
...
...
Since I have experience using this kind of AI, I know that Bing has no feelings and personality and is programmed ...
will use in the future, has no emotion, finds it useful, tool
user's emotion, future use, usefulness, emotionless
Emotional Experience

Figure 1: Thematic analysis Manual Approach

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

- We used a **middle of the road** approach (Figure 1)
- While the generated codes were based on the keywords, a continuous reiteration of the responses was conducted, allowing for the emergence of new insights through a more intuitive analysis.
- Passes throughout the dataset were not linear, but reiterative.

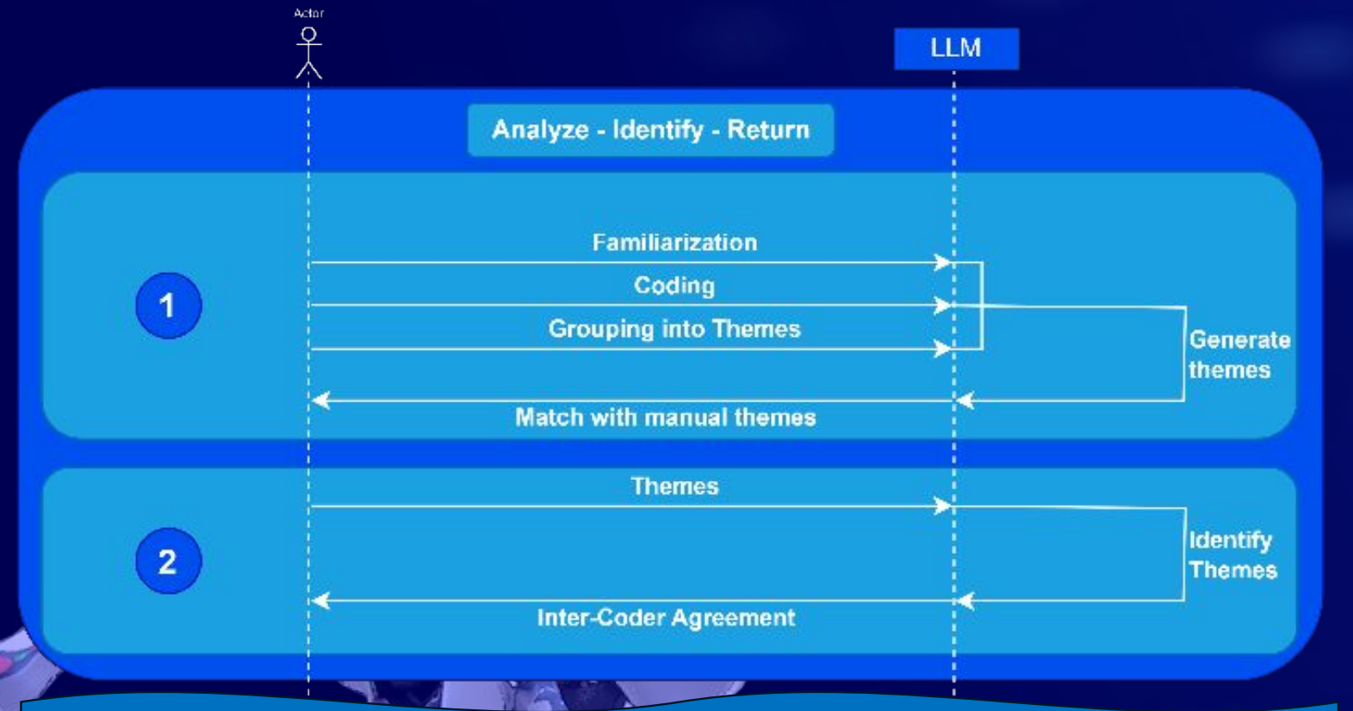


Figure 2: Large Language Model Approaches:
(1) Theme Generation Approach. (2) Theme Application Approach
Themes fed into LLMs [Qwen | DeepSeek | Phi | NeMo | Gemma | Llama]

To answer if a Large Language Model can do our task of thematic analysis, we have tried out two approaches (Figure 2).

- **Unguided Prompt**: LLM generates themes independently from the dataset
- **Guided Prompt**: LLM tags responses using predefined themes

Prompts were given following an **Analyze - Identify - Return** structure.

- Familiarize with the responses (**Analyze**), give a coding scheme (**Identify**) and group the codes together (**Return**).

Celal Karakoç

Inter-Coder Agreements

The dataset consisted of 666 open-ended responses and questionnaire scores (90 items, Likert Scale) from the ASAQ (Artificial Social Agent Questionnaire), a standardized tool for measuring user experiences with Artificial Social Agents.

- A Thematic Analysis was performed to gather themes and a **mind map** containing the themes and codes was made in between passes and updated accordingly (Figure 3).

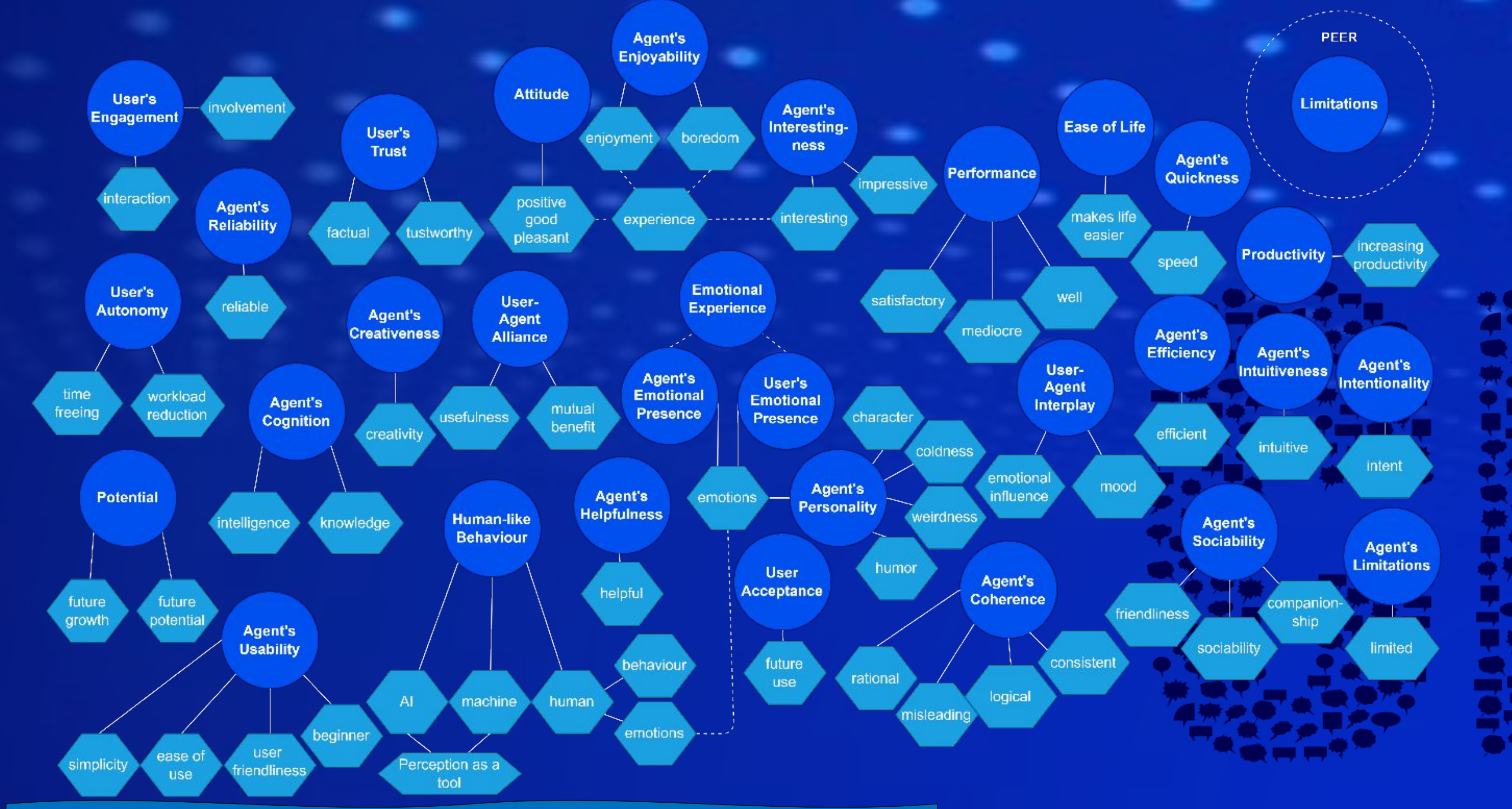


Figure 3: Mind map with regards to the *codes* (in light blue) and the resulting *themes* (in dark blue).

To **minimize bias** we did 2 Inter-Coder Agreements for each research question.

- The first Inter-Coder Agreement (Table 1) resulted in high agreements across 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.
- In cases where agreement between coders was lower, it was because the first coder identified themes that the second coder did not.

Theme (Coder 1)	Theme (Coder 2)	κ	Interpretation κ
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

Table 1: Comparison of themes with those derived by a peer (Coder 2), based on a sample of n=100 responses.

For the second Inter-Coder Agreement a mapping of the themes with the 90 items of the ASAQ was conducted by both coders. Both mappings had an overlap of 80.19%, suggesting the mapping was reliable with minimal bias.

The same process in Table 1 was conducted, but with the LLM as Coder 2. the results indicate that the LLM performs poorly when applying the coding scheme through a **guided prompt**.

As shown in Table 2 and Figure 4, the average κ over all themes is very low.

The ability of Large Language Models to reliably apply predefined themes to a set of responses remains consistently low across models.

This pattern also holds true when looking at each theme individually, rather than comparing across models.

Model	Avg. κ	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

Table 2: Average kappa across LLMs

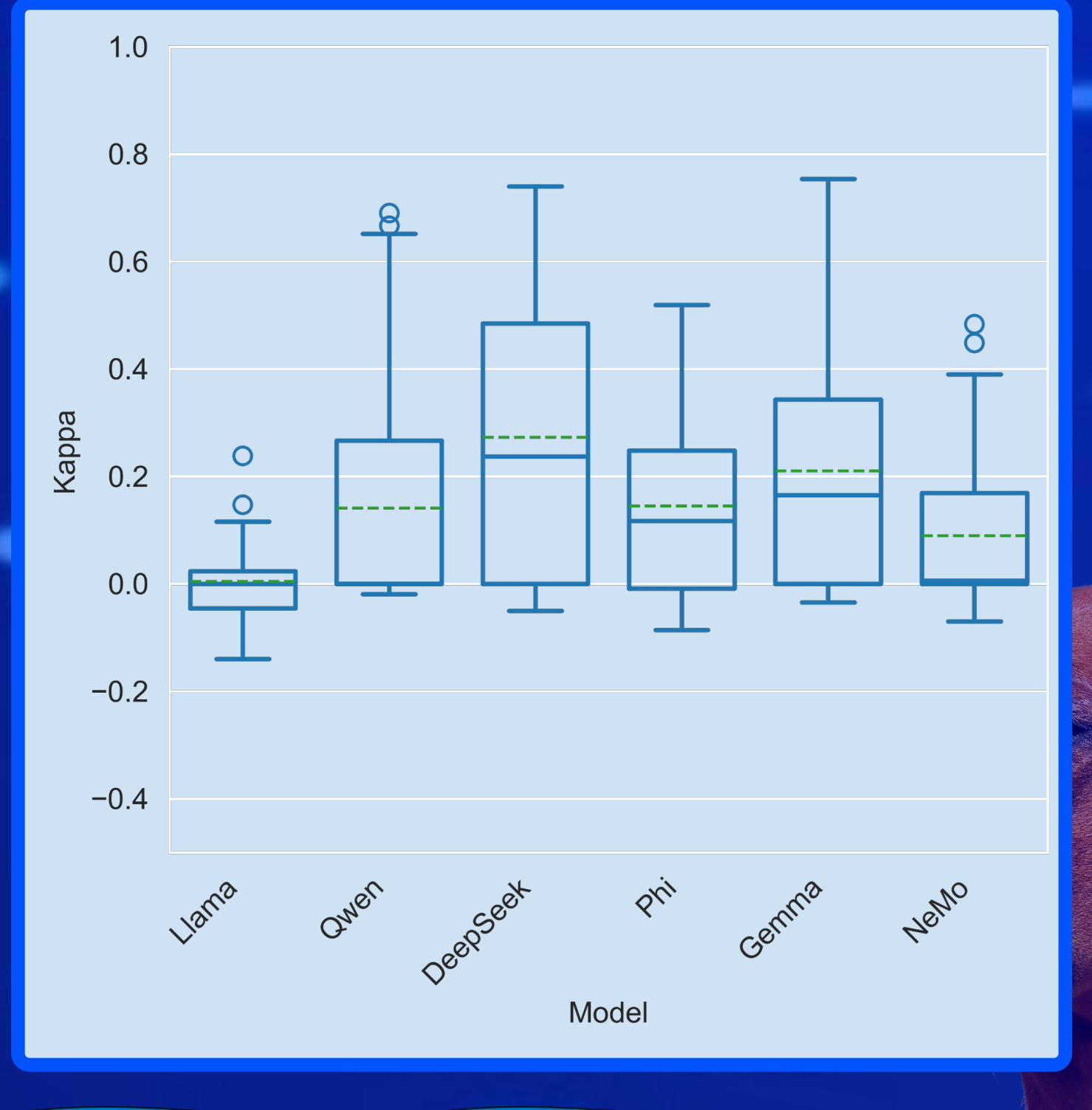


Figure 4: Kappa distributions across LLMs

Results

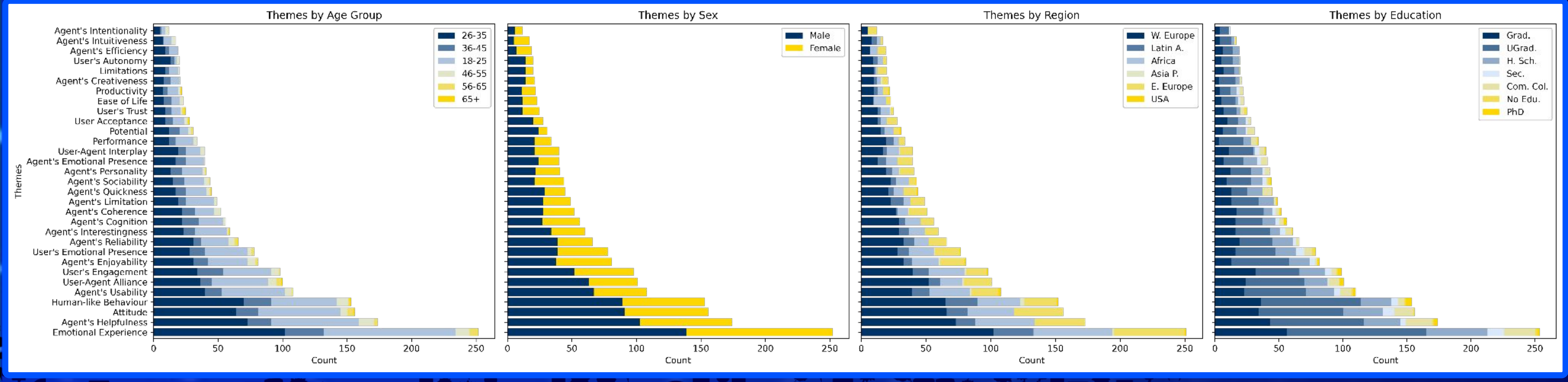


Figure 5: All the themes that were found divided into their respective descriptors.

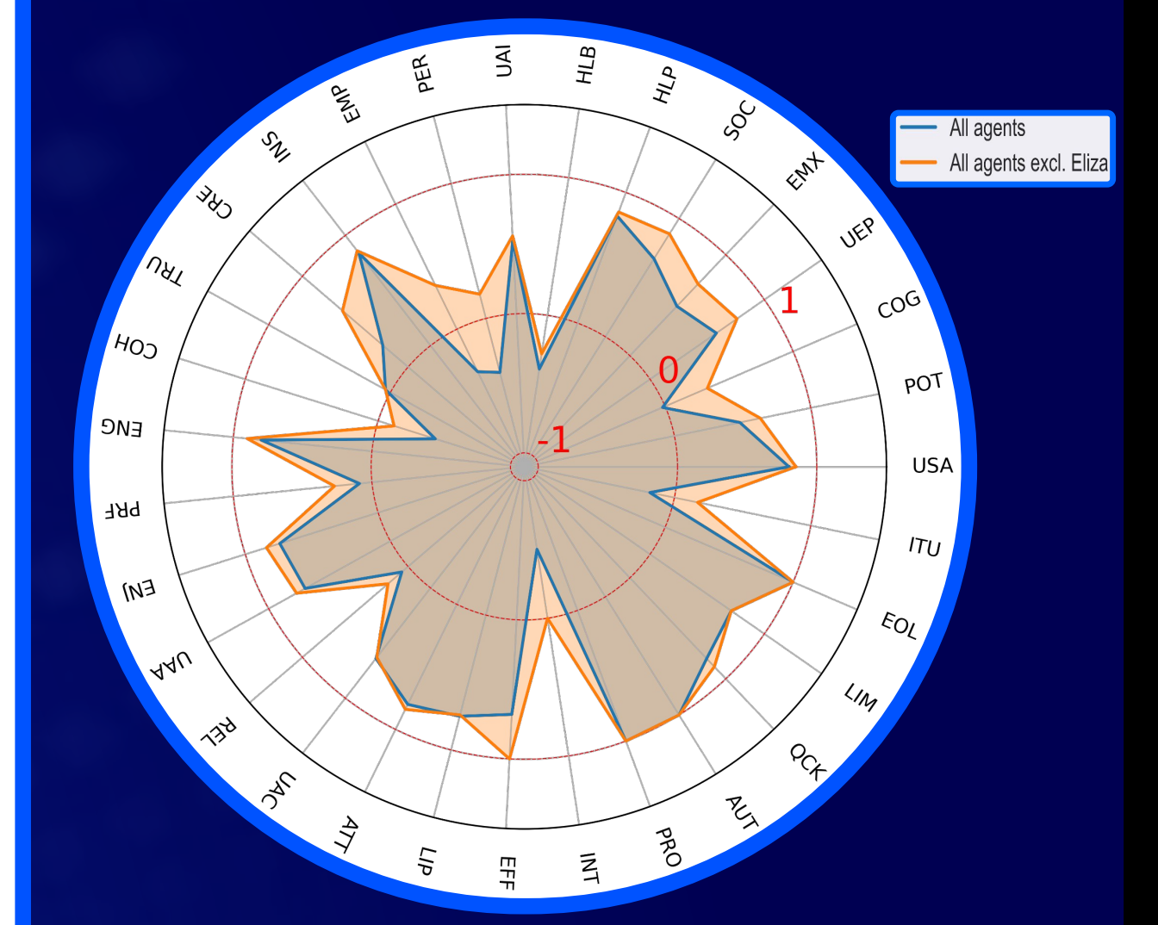


Figure 6: Average polarity (direction) of participant themes with 1 being positive and -1 being negative.

From Figure 5 we see that

- The themes throughout were quite evenly divided from their descriptors.
- There was an under representation of older age-groups
- There was an under representation of people with no formal education

Emotional Experience (with its subsets ***User's Emotional Presence*** and ***Agent's Emotional Presence***) was the most occuring theme. The themes that mattered the most towards the experience of people were the ***Agent's Helpfulness***, ***Attitude*** and ***Human-like Behaviour***. The prominence of these themes suggests participants prioritize emotional connection and practical utility in interactions with ASAs. ***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.

The agents were overwhelmingly evaluated positively, with Eliza, an early rule-based chatbot simulating a psychotherapist, standing out as an exception (Figure 6).

In regards to the theme ***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. This didn't really change how the participants felt about their usability, helpfulness, coherence or other such factors.

Theme	ρ	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 ($|\rho| < 0.1$)

Table 3: Correlations with the ASAQ

Our findings demonstrate that the ASAQ is a reliable and valid instrument for assessing user experience with Artificial Social Agents (ASAs). A substantial 74% (23 out of 31) of themes from our qualitative analysis could be directly mapped to ASAQ constructs, and 15 of these showed statistically significant, mostly positive correlations.

- Correlation analysis with the ASAQ (Table 3) showed that all statistically significant themes had positive correlations with their ASAQ counterparts, which suggests that higher user ratings on aspects like enjoyability, intelligence, and reliability align with higher overall ASAQ scores.
- Themes such as ***Agent's Cognition***, ***Enjoyability***, ***Reliability***, and ***Sociability*** had especially strong correlation.
- Two unexpected results are noticeable: ***Agent's Quickness***, which might lack a clear ASAQ counterpart and may not align with any ASAQ construct, and ***User's Engagement***, which showed a weak negative correlation. Any conclusions based on the ***User's Engagement*** theme should be interpreted with caution.

Conclusions

This study explored user experiences with Artificial Social Agents (ASAs) through manual thematic analysis. 31 distinct themes were identified.
We demonstrated that

- Key themes including ***Agent's Helpfulness***, ***Attitude***, and ***Human-like Behaviour***, highlight users' desire for both practical support and emotional resonance.
- 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.
- ***Agent's Cognition***, ***Coherence***, and ***Intentionality*** show that users assess intelligence of the agent with a mixed sentiment
  - people praised insightful responses but quickly noticed incoherence, contradictions or awkward behaviour.
- 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 correlated with a desire to continue using the agent.
- **Tone mattered**: cold or mechanical interactions led to negative sentiment, while warm or natural responses improved user perception

In addition to the manual analysis, we evaluated the capabilities of various locally hosted Large Language Models (LLMs) in conducting thematic analysis.
We demonstrated that

- LLMs were capable of capturing general thematic structures through ***unguided*** prompts (achieving a 74% overlap with manually identified themes)
  - The high overlap suggests 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.
- LLMs performed poorly in the ***guided***, response-level thematic annotation task.
  - i.e. LLMs lack the consistency and nuance required for fine-grained qualitative analysis, particularly when applying predefined coding frameworks.

References

[1] - github.com/ckarakoc/bep-asa (**DATA & CODE**)
[2] - ii.tudelft.nl/evalquest/web (**ASAQ**)

TU Delft