Comparing Local LLM-Based Extraction of Stakeholder Values for Value Model Visualization in Deliberations

How effectively can local large language models extract information required to generate visualizations of stakeholder value models?

Maximiliaan van der Veek[1]
Supervisor(s): Willem-Paul Brinkman[1], **Michaël Grauwde**[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 21, 2026

Name of the student: Maximiliaan van der Veek
Final project course: CSE3000 Research Project
Thesis committee: Willem-Paul Brinkman, Michaël Grauwde, Sole Pera

Abstract

In stakeholder deliberation, it can be useful to give insights into stakeholders' values (such as *privacy*, *safety*, and *fairness*). Previously, conversational agents were built to give insight into such a value model. Yet, the question of whether visualizations can aid in the understanding of a value model still remains. As a first step, this paper proposes two visualizations of value models based on existing literature: radar charts and value cards. This paper argues why these visualizations can aid in the understanding of value models, but to create them, data needs to be extracted from transcripts. Therefore, we also compare three consumer-grade local LLMs ($\leq$ 35B parameters) – Gemma4:e4b, Phi4-reasoning:14b, and Qwen3.6:35b – on their ability to extract data from deliberative transcripts necessary to generate these visualizations. Using local LLMs for this task can be beneficial, as using cloud-provided LLMs can lead to value profiles being built. The evaluated LLMs are found to have strong agreement (Cohen's $\kappa \geq 0.808$) with human coders on extracting the values included in the transcripts; however, they have mixed agreement when ranking the values or assigning codes to their importance. When comparing textual justifications for value rankings and assigned importance codes, justifications between LLMs and humans may differ, but the textual justifications generally do not disagree on whether a value is important or not. When the local LLMs have to give a summary of the meaning of values, they are generally *roughly* similar[1] to human summaries or those provided by the other LLMs, and in 38 to 46% of cases, LLM summaries are nearly equivalent.[1] These findings suggest that consumer-grade local LLMs are effective at identifying human values present in text, but struggle to code their importance, ordinal rank, and summarize their meaning. This makes them currently unsuitable to replace human coders without oversight.

1 Introduction

Conversational agents (CAs) have become very prominent in several domains, with examples ranging from cognitive behavior therapy (Woebot) [15] to customer service on social media [47]. This proliferation is reflected in the market size of chatbots, which is projected to grow from 7.76 billion USD in 2024 to 27.29 billion USD by 2030 [13]. As conversational agents become increasingly integrated in our social fabric, researchers are also examining their role in democratic discourse [22]. One aspect of democratic discourse is the deliberation process. Declining trust in democracy in many countries makes this a hot topic [42]. Traditionally, the deliberation process aims for consensus [18, 43]. However, the aim of reaching consensus has been criticized for leaving the confrontational nature of democratic discourse out of the equation, with a risk that the focus can become one of conformity, excluding participants' views and ideals from the discussion [8, 18, 37, 43]. As Kerkhof puts it: "Whereas consensus building can be characterized as a process of negotiation, deliberation is about dialogue and argumentation" [43, p. 282].

In stakeholder dialogue, facilitating the disclosure of stakeholder interests and motives can lead to a better understanding of the problem at hand and increase awareness of participants' underlying assumptions [43]. It can also affect stakeholder preferences and lead to participants developing a mutual understanding of each other's values [18, 21, 33]. One of many ways to reveal underlying assumptions in stakeholder dialogue is value-focused thinking [23]. In value-focused thinking, values, rather than alternatives, should be the "driving force" of the decision-making process. The rationale behind this is that because the available options are merely mechanisms to realize objectives, identifying what one truly cares about should precede the evaluation of choices.

[1] For further details, we refer the reader to Section 4.2.4 and Appendix G.

Research on chatbot-moderated deliberation in general employs agents designed to foster consensus [18, 26]. However, relatively little research looks into the general design of conversational agents with a focus on value explication and reflection [18]. On top of that, prior research did not examine whether visualizations can aid in the value explication or reflection process. Visualizations can be relevant for value explication because, if done well, they can amplify cognition in several ways. One such way is reducing cognitive load by offloading work from the cognitive to the perceptual system [5]. Additionally, illustrated text can improve text comprehension [17]. Therefore, this paper first suggests two visualization methods for stakeholder value models. A value model here refers to a collection of (human) values that arise in a deliberative transcript. Because these visualizations rely on data being extracted from deliberative transcripts, this paper also investigates whether consumer-grade local LLMs can be used for this data extraction. This can be useful, as human extraction of this data can be time-intensive. For the purposes of this research, (consumer-grade) local LLMs are defined as models with 35 billion parameters or fewer ($\leq$ 35B), as they can be executed effectively on standard consumer hardware.

The main research question is:

> *How effectively can local, open-source large language models extract the information required to generate visualizations of stakeholder value models?*

Subquestions:

> *(1) What design heuristics or properties are desirable for visualizing stakeholder value models in a deliberative context?*
>
> *(2) How consistent are local LLMs with human ground-truth and each other when evaluating value inclusion, meaning, rankings, importance, and their corresponding textual justifications?*

The rest of this paper is structured as follows: The relevant findings to answer the first subquestion are described in Section 2.1. The rest of Section 2 describes the theoretical basis for using LLMs to extract human values from transcripts. In Section 3, the methods used to answer the research questions are described. Both visualizations are shown here, and the process of comparing LLMs to human judgments is described. Then, in Section 4, the results are described. Section 5 describes the considerations made for doing this research responsibly. The paper ends with a discussion, suggestions for future work, limitations, and conclusions.

2 Related Work

2.1 Desirable properties of visualizations

Not many papers study the specific desirable properties of human value visualizations. Existing visualizations for human values (e.g. [11, 12, 25]) usually do not academically test or validate the imagery used [19]. This section looks into more detail on how exactly (in general) visualizations can contribute to the comprehension of value models, and what properties of such visualizations are desirable.

2.1.1 How visualizations aid cognition

Larkin and Simon [30] illustrate how diagrammatic visualizations can aid cognition. They concluded that visualizations help by: (1) grouping relevant information together, (2) using location to group information about a single element, and (3) automatically supporting a large number of perceptual inferences that are extremely easy for humans. This conclusion is further backed by Card, Mackinlay and Shneiderman [5]. They propose six ways in which visualizations can amplify cognition (p. 16). For the purposes of this research, the most important of these six properties are: (1) increasing the memory and processing resources available to users, (2) reducing the search for information, and (3) enabling perceptual inference operations. Schematization in diagrams can also reduce the amount of irrelevant information, allowing subjects to focus on important features, increasing speed and accuracy of information processing [40].

2.1.2 Sensory and arbitrary codes

When using symbols in visualizations or illustrations, the choice of the type of symbol can affect the effectiveness of the visualization. An important distinction in visual symbols is that between sensory and arbitrary symbols [44]. Sensory aspects of visualizations are defined as those that derive expressive power by using the perceptual processing of the brain, without learning. Arbitrary symbols are defined as conventions that derive their power from culture.[2]

Several important properties of sensory symbols (or codes) include: (1) sensory codes can convey meaning without any additional training, (2) sensory codes are resistant to instructional bias,[3] and (3) sensory codes will, in general, transcend cultural boundaries.

In contrast, "arbitrary codes are by definition socially constructed" [44, p. 15]. They are culture-specific.[4]

2.1.3 Desirable design properties, specific to human value visualizations

To find more desirable design properties of visualizations specific to human values, Harms' proposed set of illustrated cards [19] can aid. In her research, these cards were found to be the most contributing to participants' understanding of the human values. It is important to note, however, that her solution only contains five values, whereas the dataset in this research contains eight. Harms found that visualizations of human values should contain the following four elements: (1) the name of the value, (2) a description of the value, (3) an illustration (image) of the value, and (4) written context. This is consistent with other state-of-the-art visualizations in the form of cards (e.g. [6,11,24]) that use words and images together. This use of words and images together is also backed by Tufte [39].

[2] An example of a sensory symbol would be a cave painting of humans hunting a deer, whereas the written word dog would be an arbitrary symbol. It derives its meaning from the fact that everyone agrees on its meaning. It could just as well have been another word, like platypus, but people have agreed on a different meaning for that word.

[3] As an example, many illusions still work despite knowing that they are illusions, as illustrated on page 14 of [44].

[4] An example described in the book (p. 16) is that many geologists still use a topographic contour map, rather than shaded computer graphics representations, even though the shaded representation can be more intuitive for most people. Another example (p. 16) illustrates how the color green is often used in applications to signify a correct or successful action and red for warnings, while in Chinese culture, green symbolizes death, and red symbolizes luck and good fortune.

2.1.4 Important properties of visualizations specific to this research

Based on the literature, several important properties that the visualizations in this research should have were found. They are summarized in Table 1.

Table 1: Important properties visualizations in this research should have, based on the literature.

Source(s)	Property
[5, 30, 40]	Group relevant information together (thereby reducing the search for information).
[5]	Show a relation between the different items (values).
[5, 30]	Enable participants to perform perceptual inference operations.
[44]	Use sensory symbols (as these can transcend cultural boundaries and they can be universally understood).
[44]	When using arbitrary codes or diagrammatic representations, convention is key to reducing the learning curve. Widely used representations are preferred over more niche ones.
[44]	Avoid color codes to convey meaning, as the meaning of colors is highly culturally specific.
[19]	Ideally, visualizations of human values should contain four elements: (1) the name of the value, (2) a description of the value, (3) an illustration (image) of the value, and (4) written context.
[19]	Illustrations of human values should not contain any texture, pattern, shadow, or background, and people in illustrations should have facial details. The illustrations should be neither too simple nor too overwhelming.

2.2 LLMs and Human Values

As outlined in Section 1, this research is not only interested in which values are present in a transcript, but also in how those values relate to one another. Prior research has, however, mostly focused on the former. Brigola [4] evaluates different LLMs on their ability to identify the values present in a text, finding no significant difference between LLMs and noting that adding a description to each value improves identification accuracy. They use LLMs with 7 to 8 billion parameters, which can be run locally. Zhu et al. [50] extend this by adding an online LLM into the picture. They use a small local LLM to generate initial estimations used to optimize prompts for a larger, online LLM, reducing its token usage and improving its accuracy. Both studies, however, only focus on using LLMs to identify which values are present in a text. To distinguish this from also extracting other relevant data, the rest of this paper refers to the identification of values present as "value inclusion."

Other works also include importance ("intensity") ratings. Kim et al. [27] compare several LLMs against humans on rating each value's intensity from 1 to 10, finding mixed results that vary depending on the LLM used. De la Cruz et al. [10] similarly add an intensity-coding task and propose a model architecture for it, but their best-performing model only achieves an F1 score of 0.34 against human judgments. These studies, however, still treat values independently. None of the studies discussed extract an ordinal ranking, which arguably requires higher-level reasoning as the LLM must retain the entire set of values in its active context window and compare them. They also do not extract the textual justifications or contextual summaries of these values. This research extends prior work along all of these dimensions, extracting not only value inclusion but also ranking, importance codes, justifications for rank and importance, and a summary of each value's meaning.

3 Method

A dataset collected in currently under-review work by Grauwde was used.[5] It contains 20 transcripts of conversations between a human participant and an LLM. In these conversations, the LLM tries to make the participant reflect on their own values in a public safety decision-making context. The LLM then tries to build a model of their values based on the participants' responses. This makes the dataset very applicable to this research, as the subject of the conversation is always about stakeholder values in a decision-making context.

3.1 Visualizations of stakeholder value models

Based on the literature described in Section 2.1, two visualization methods were selected: radar charts and value cards. The code to generate both visualizations based on LLM JSON output is provided on GitHub.[6]

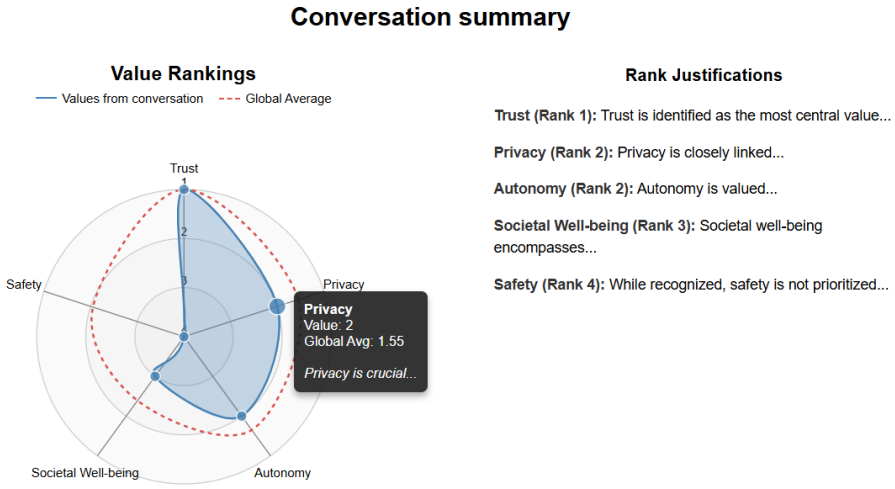

Figure 1: Radar-chart visualization of a stakeholder value model. Each value is one axis. It shows the global average as a dotted red line, and the stakeholder model in blue. On the right side, explanations as to how the ranking was determined are provided. When the stakeholder hovers over a blue dot, a tooltip with a summary is shown.

Radar charts were chosen because they allow to effectively group the values and their rankings in a single figure, reducing the need to search. Additionally, they allow for fast perceptual inference operations because they allow the participant to visually distinguish relative rankings or importance assigned to each value. They also allow participants to compare their value model to the average model of the group. This is an inference that could not be made solely based on the transcript itself, as the transcript does not provide this information.

An example visualization using radar charts is given in Figure 1. The radar chart is modified to have smooth curves, as objects with smooth curves are easier for humans to perceive than those with abrupt changes [44]. A blue/red color scheme is used to differentiate between data related to values from the conversation and the average model from the group,

[5] https://data.4tu.nl/private_datasets/xUIgPm6lCdMpkQef_C-AaPCswJzKgqCXdBhq2nFVIJY.

[6] https://maximiliaanvdv.github.io/CSE3000_LLM_Extraction_Stakeholder_Values/downloads/LLM_Data_Extraction_Supplementary_Files.zip

as this color scheme is colorblind-friendly [31, 32, 34]. To also reduce the dependency on color alone to make this distinction (as recommended by [9, 32, 41]), the values from the conversation are shown as a solid line, and the average model is a dotted line. This allows them to be distinguished by shape, instead of color alone.

Radar charts are diagrammatic representations, yet many state-of-the-art visualizations are in the form of cards (e.g. [6,11,12,24]). By representing the values as cards, an illustration of the value itself can be provided as well. This provides a sensory code to the value. The value description and written context can also be provided on the card, reducing the need to search even more than in the radar-chart representation of the value model. This comes at the cost of not being able to quickly compare value importances. An example value-card visualization is provided in Figure 2.

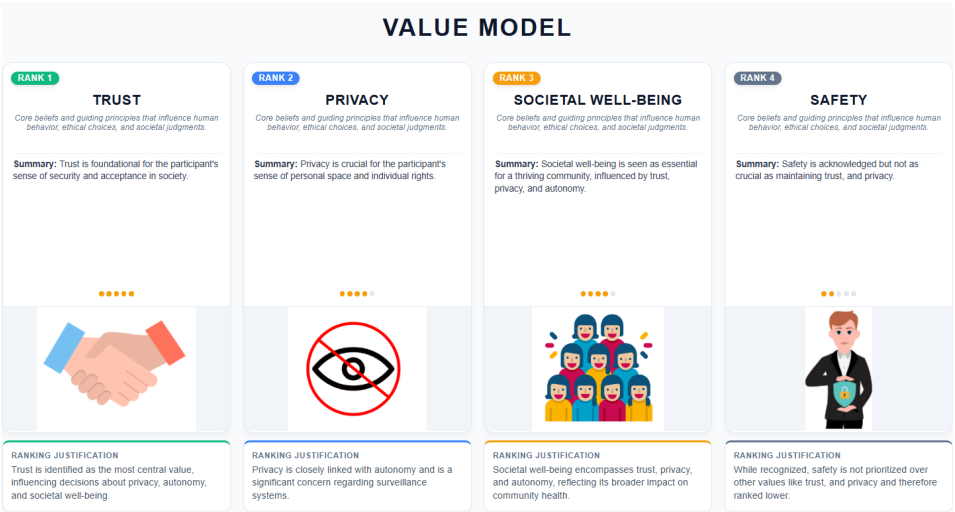

Figure 2: Value-card visualization of a stakeholder value model. Each value is one card. Each card shows the rank in the top-left. The card's title is the value name. Below that is a general definition of the value. The summary says what this value means in the context of the deliberative transcript from which this value is taken. The orange dots above the image show the importance for the participant (more dots meaning higher importance). Below each card is a justification as to why the card is assigned its rank. The accent colors green, blue, orange, and grey are purely to associate the correct justification to the correct card, but do not have any further meaning.

3.2 Extracting value models from transcripts

To answer the second sub-research question, data needed to be extracted from each of the 20 transcripts. In particular, five aspects needed to be extracted in order to create the visualizations proposed in Section 3.1: (1) a numeric, ordinal ranking of each value that exists in a deliberative conversation or transcript, (2) a numeric code from 1 to 5, assigned to the importance of each value (where 1 is very unimportant, and 5 very important), (3) a textual justification for the rank assigned to that value, (4) a textual justification for the importance code assigned to each value, and (5) a summary of what the value means in the context of the conversation or transcript.

Items 1, 2, and 5 are absolutely necessary. Without them, the information in the visualizations is incomplete. Textual justifications (items 3 and 4) are added to illustrate the integrity of the content displayed. Clarifying the origins of the data in a visualization is a

principle recommended by Tufte [38]. It allows the viewer to see and critically assess what the data in items 1, 2, and 5 is based on.

Two coders (the author of this research and another researcher familiar with the values) coded each conversation. They first independently coded each conversation according to the instructions found in Appendix A. Based on this coding, eight values were identified across all conversations (autonomy, privacy, safety, societal wellbeing, transparency, accountability, fairness, and trust). It is important to note that value models were extracted from each conversation in its entirety, and not from each individual statement. Multiple statistical measures were calculated to check if the coding was reliable. The independent coding, final coding, and spreadsheets used for all calculations relating to human and LLM evaluations can be downloaded from GitHub.[6]

3.2.1 Inter-coder reliability for value inclusion

To check whether both coders agreed on *the values included in each transcript*, the rankings were transformed into Boolean data. Per conversation, any value included in the ranking was given a 1, and a 0 if it was not ranked. Then, Cohen's κ was calculated for the transformed data. This statistic was chosen because (unlike the overall agreement rate) it also accounts for the agreement that occurs based on chance alone. The overall percentage of agreement for value inclusion was 96.9. Cohen's κ was 0.924, showing an almost perfect agreement according to the interpretation table by Landis and Koch [29].

3.2.2 Inter-coder reliability for value ranking

To see if both coders agreed on the *rankings* of each value, Kendall's τ_b test (with ties) was used, together with Spearman's ρ and a weighted Cohen's κ.[7] Kendall's τ_b test is suitable for this data, since it measures correlation and does not assume a normal distribution. Additionally, the data has ordinal scale levels. The same reasoning holds for Spearman's ρ. It is important to note, however, that both Spearman's ρ and Kendall's τ_b do not measure the degree of disagreement. To measure the degree of disagreement, a *weighted* Cohen's κ was also calculated. The weights used were the absolute differences of the ranks assigned by each coder. This penalizes larger disagreements more than smaller ones by a linear factor.

When including values ranked as 'not present' by both coders, Kendall's τ_b was 0.868, Spearman's ρ 0.910, and the weighted Cohen's κ 0.874. This high agreement can occur if every transcript is assumed to have the set of all unique values found across all transcripts, since the agreement on value inclusion is very high. In this case, many values are agreed to be 'not present', and are thus ranked the same. Therefore, we also calculate these statistics separately, excluding the values that were marked 'not present' by both coders. When excluding these values, Kendall's τ_b was 0.745, Spearman's ρ 0.796, and the weighted κ 0.820. Kendall's τ_b and Spearman's ρ show very high correlation in the value rankings, and Cohen's κ shows almost perfect agreement.[8]

[7] This was chosen over the Krippendorff-α statistic for ordinal data, because the same values appear in several conversations, and they do not necessarily have the same rank in each conversation. This leads to a low value of the Krippendorff-α (0.65), despite high overall agreement.

[8] Kendall's τ_b values were converted to Spearman's ρ using [16], then interpreted according to [28, p. 195]. Spearman's ρ interpreted according to [28, p. 195]. Cohen's κ interpreted according to the interpretation table by Landis and Koch [29].

3.2.3 Inter-coder reliability for value importance

To find out if both coders agreed on the *importance* of each value, the data was first transformed by assigning each label of importance a number:
Unranked = 0, Not important = 1, (Somewhat) unimportant = 2, Neutral = 3, (Somewhat) important = 4, Very important = 5.
Then, Kendall's τ_b, Spearman's ρ, and the weighted Cohen's κ were calculated on the importance values as well. When including the values that were coded as 'not present' by both coders, Kendall's τ_b was 0.875, Spearman's ρ 0.910, and the weighted κ 0.861. When excluding these values, Kendall's τ_b was 0.777, Spearman's ρ was 0.820, and the weighted κ 0.860. Kendall's τ_b and Spearman's ρ show very strong correlation and the weighted Cohen's κ shows almost perfect agreement in the importance assigned to each value by the coders.

3.2.4 Triangulating the human transcript evaluations

After coding the values and determining that the coding was reliable, both coders discussed each value that was coded differently to reach a final coding for each transcript. Both coders eventually reached consensus on the coding of all values. In certain cases, values were omitted or grouped under another value. An example of this is conversation R3P2, where coder A included *transparency* as a separate value, whereas coder B argued the participant gave insufficient reasoning to code it as a separate value, relating the transparency aspect back to privacy. Since coder A agreed with this reasoning, in the final coding, transparency is grouped under privacy. The justifications for each value ranking and value importance were triangulated by the coders in a similar fashion as well.

3.2.5 Using LLMs to code the transcripts

After triangulating the human evaluations, three open-source LLMs were tasked to create a similar evaluation of each transcript: Qwen3.6:35b,[9] Phi4-reasoning:14b,[10] and Gemma4:e4b.[11] Cloud-models were not considered in order to make reproduction of this research more accessible and free. Another reason for not considering cloud-models is the potential for data leaking to cloud providers, which could be used to build value profiles. This is further explained in Section 5. Qwen3.6:35b was chosen because of its high score (86%) on the GPQA Diamond benchmark [35, 36], showing its reasoning qualities. Phi4-reasoning:14b is smaller than Qwen3.6:35b with only 14 billion parameters. Despite this, it still scores relatively high (67.1%) on the GPQA Diamond benchmark, rivaling models like DeepSeek-R1 [1]. Gemma4:e4b is the smallest model of the three chosen. It performs the worst out of the three on the GPQA Diamond benchmark, scoring just 58.6% [14]. However, it is still included to also test a model with a relatively small number of parameters. The three models were tasked to extract the same data from each transcript. The system prompt used is provided in Appendix B. This is a chain-of-thought (CoT) prompt. It asks the LLM to think in a structured fashion and provide its reasoning along the way. This is useful, since this is a reasoning task, and in general, CoT prompting elicits reasoning in large language models [45]. Zero-shot CoT (providing only the task description) was used instead of few-shot CoT (providing task description and some examples), since the size of the dataset was limited. However, it is important to note that CoT prompting may not work fully effectively with these models, as they each have less than 100 billion parameters [45]. The same sys-

[9] https://ollama.com/library/qwen3.6
[10] https://ollama.com/library/phi4-reasoning
[11] https://ollama.com/library/gemma4

tem prompt (Appendix B) was used for all models. The process of establishing this system prompt is described in Appendix C. Across all tests, the parameters were kept the same (temperature = 0, seed = 42, top_k = 1, top_p = 1). Each transcript was provided to the LLM in textual format using the code found in Appendix D.

3.2.6 Comparing textual justifications between LLMs and human transcript evaluations

The human coders and the LLMs created textual justifications for the value rankings and the importance assigned to each value. To compare their similarity, each possible justification pair was coded by both human coders separately. A score of 2 was assigned if both justifications either cited the same quotes from the transcript, or described the same reason for giving a certain rank or importance code to the value. A score of 1 was assigned if different citations were given or described for providing a rank or importance code to the value. Finally, a score of 0 was given to justification pairs that had active disagreement. An example of active disagreement is a pair of justifications where one provides a reason why the value is unimportant, while the other gives a reason why the value is important. Example justification pairs with codes are given in Table 7, Appendix E.

Similarity scores could sometimes be automatically assigned. A 0 was automatically assigned to justification pairs where one coder excluded the value, while the other did not. A 2 was automatically assigned when both coders excluded the value. Automatically assigned scores were skipped by the human coders to save time. As this could lead to inflated reliability statistics, they are also calculated excluding automatically coded scores.

The inter-coder reliability of the assigned similarity scores was evaluated using a weighted Cohen κ, using the absolute difference between assigned similarity scores as weights. This statistic is again chosen because the data is ordinal, and it measures the degree of disagreement while accounting for the chance of agreement. The weighted Cohen κ was calculated for each LLM-LLM and LLM-human justification pair. The lowest Cohen κ, being roughly 0.851, shows almost perfect agreement between both coders.[12] All values for Cohen κ are given in Appendix F, Table 8.

3.2.7 Comparing value summaries

All the values that had a rank and importance assigned were also given a textual summary of the value's meaning within the context of each transcript. The similarity of each pair of summaries provided was evaluated on a scale from 0 to 5. A score of 0 was assigned to a pair of sentences that were completely dissimilar, and a 5 was assigned when two sentences in a pair were equivalent. The full, detailed instructions can be found in Appendix G. These instructions are adapted from [7], which is a task designed to assess the semantic similarity of sentence pairs. Similarly to the rank and importance justifications, a 0 was automatically assigned to pairs where one coder excluded the value and the other did not, and a 5 when both excluded the value. Inter-coder reliability was again evaluated using a weighted Cohen κ using the absolute difference between assigned scores as weights. The weighted κ is calculated separately for each combination of coders that created the summaries. When including automatically coded scores, the lowest κ was 0.863, showing almost perfect agreement. When excluding the automatically coded scores, the lowest κ was 0.613, which still shows substantial agreement. Spearman's ρ and Kendall's τ_b were also calculated, showing high to very high correlation. These values can be found in Appendix H, Table 10.

[12] Interpreted according to Landis and Koch [29].

4 Results

4.1 Value model visualizations

Based on the literature described in Section 2, a radar-chart and value-card visualization were proposed, in order to visualize human value models that can arise from a deliberative transcript. These visualizations are shown in Figures 1 and 2, respectively. The code to generate them can be found on GitHub.[6]

4.2 Data extraction comparison

In order to create the visualizations, data needs to be extracted from transcripts. The remainder of this subsection describes the comparisons between LLMs and human data extracted from deliberative transcripts.

4.2.1 Value inclusion

Overall, all three LLMs had an agreement rate ranging from substantial to almost perfect on value inclusion when compared to human judgments. Gemma4 scored highest in both cases (Cohen's κ of 0.896 when including values marked as 'not present' by both[13] coders, 0.848 when excluding them), while Qwen3.6 scored the lowest (Cohen's κ of 0.870 when including values marked as 'not present' by both coders, 0.808 when excluding them). Appendix I, Table 11 shows Cohen κ values achieved by the LLMs compared to human judgments.

When comparing pairs of LLMs, all pairs had an almost perfect agreement rate. Overall, Phi4-reasoning paired with Gemma4 had the lowest agreement (Cohen's κ of 0.921 when including values marked as 'not present' by both coders, and 0.887 when excluding them). Qwen3.6 and Phi4-reasoning had the highest pairwise agreement (Cohen's κ of 0.948 when including values marked as 'not present' by both coders, and 0.924 when excluding them). All κ values achieved by pairs of LLMs are given in Appendix I, Table 12.

4.2.2 Comparing codes assigned to value rank and importance

Comparing the value rankings provided by the LLMs against the triangulated human rankings showed medium to very high agreement. All values of ρ, τ_b, and κ per model compared to human ranks are given in Table 2. They are also given in Appendix J, Table 13.

When including values ranked as 'not present' by both coders, we can see that overall agreement is very high.[14] However, for the same reason described in Section 3.2.2, we calculate these statistics separately, excluding values that were ranked as 'not present' by both coders. When excluding the values ranked as 'not present' by both coders, the weighted Cohen's κ shows moderate agreement, Kendall's τ_b shows medium to high correlation, and Spearman's ρ shows medium correlation. These differences can be explained by the fact that the weighted Cohen's κ calculates the degree of disagreement, while Spearman's ρ and Kendall's τ_b measure the direction of paired ranks instead of the degree of difference between ranks.

The weighted κ shows moderate agreement for all models. ρ and τ_b show Phi4-reasoning had high correlation, and Qwen3.6 and Gemma4 had medium correlation. In all cases, Phi4-reasoning had the highest correlation with humans when determining the rankings of values.

[13] Both meaning: the model compared to the triangulated human codes.

[14] $\rho \geq 0.70$, $\tau_b > 0.51$, and $\kappa > 0.8$

Table 2: Spearman ρ, Kendall τ_b, and weighted Cohen κ (rounded to 3 decimals) for value ranking compared to human judgments per model. Highest values per statistic are marked in **bold**.

	Qwen3.6:35b			**Phi4-reasoning:14b**			**Gemma4:e4b**		
Statistic	ρ	τ_b	κ	ρ	τ_b	κ	ρ	τ_b	κ
Incl. values marked absent by both coders	.816	.774	.841	**.875**	**.828**	**.859**	.853	.792	.849
Excl. values marked absent by both coders	.352	.308	.536	**.590**	**.497**	**.570**	.392	.307	.553

Comparing the codes given to value importance by LLMs to that of humans again leads to very high agreement when values coded as 'not present' by both coders are included. When excluding them, moderate to high agreement is achieved. Similar to the value rankings, Phi4-reasoning had the highest agreement with human importance codes when excluding values marked as 'not present'. Phi4-reasoning and Gemma4 achieved high correlation on the τ_b statistic, but moderate on ρ and κ. Qwen3.6 achieved medium correlation on all statistics. The agreement rates for LLM-coded value importance compared to human evaluation can be found in Table 3. They are also given in Appendix J, Table 14.

Table 3: Spearman ρ, Kendall τ_b, and weighted Cohen κ (rounded to 3 decimals) for value importance compared to human judgments per model. Highest values per statistic are marked in **bold**.

	Qwen3.6:35b			**Phi4-reasoning:14b**			**Gemma4:e4b**		
Statistic	ρ	τ_b	κ	ρ	τ_b	κ	ρ	τ_b	κ
Incl. values marked absent by both coders	.822	.785	.819	**.855**	**.814**	**.839**	.851	.806	.832
Excl. values marked absent by both coders	.393	.348	.470	**.486**	**.421**	**.513**	.415	.357	.505

A pairwise comparison of the value rankings assigned by each LLM reveals very high agreement across all model pairs, when including values that both models marked 'not present'. When they are excluded, all model pairs show high correlation or substantial agreement on all statistics (ρ, τ_b, and κ). All values for ρ, τ_b, and κ are shown in Table 4, as well as in Appendix J, Table 15.

Table 4: Spearman ρ, Kendall τ_b, and weighted Cohen κ (rounded to 3 decimals) for value ranking in pairwise LLM comparison (Qwen3.6:35b shortened to "Qwen", Phi4-reasoning:14b shortened to "Phi4", Gemma4:e4b shortened to "Gemma4"). Highest values per statistic are marked in **bold**.

	Qwen, Phi4			**Qwen, Gemma4**			**Phi4, Gemma4**		
Statistic	ρ	τ_b	κ	ρ	τ_b	κ	ρ	τ_b	κ
Incl. values marked absent by both coders	**.914**	**.867**	**.909**	.908	.861	.903	.891	.832	.881
Excl. values marked absent by both coders	.562	.505	**.704**	**.583**	**.510**	.699	.507	.434	.623

When comparing the value importance labels assigned to each value by pairs of LLMs, correlation is also high. When excluding values coded as 'not present' by both models, all pairs achieve high correlation for Spearman's ρ. Phi4-reasoning paired with Gemma4 achieve high correlation on the Kendall's τ_b statistic. The other pairs achieved very high

correlation on Kendall's τ_b. All pairs achieved substantial agreement on Cohen's κ, except Phi4-reasoning paired with Gemma4 which achieve moderate agreement. In all cases, Qwen3.6 paired with Phi4 achieved the highest ρ (0.647), τ_b (0.590), and κ (0.695). Table 5, and Appendix J, Table 16 show the values of ρ, τ_b, and κ for all model pairs.

Table 5: Spearman ρ, Kendall τ_b, and weighted Cohen κ (rounded to 3 decimals) for value importance in pairwise LLM comparison (Qwen3.6:35b shortened to "Qwen", Phi4-reasoning:14b shortened to "Phi4", Gemma4:e4b shortened to "Gemma4"). Highest values per statistic are marked in **bold**.

	Qwen, Phi4			**Qwen, Gemma4**			**Phi4, Gemma4**		
Statistic	ρ	τ_b	κ	ρ	τ_b	κ	ρ	τ_b	κ
Incl. values marked absent by both coders	**.922**	**.886**	**.907**	.914	.878	.893	.883	.838	.869
Excl. values marked absent by both coders	**.647**	**.590**	**.695**	.657	.584	.664	.515	.451	.591

4.2.3 Comparing the similarity of textual justifications for rank and importance

When comparing the textual justifications for both rank and importance given by LLMs against those given by humans, in over 90% of all cases, a similarity score of 1 or 2 is given. When comparing pairs of justifications given by LLMs, a similarity score of 1 or 2 was assigned in over 95% of cases. This means that in over 90% of cases, while the justifications may refer to different statements from a transcript (score 1), LLMs do not argue that values are unimportant when the human deemed them important (score 0). Further statistics relating to the assigned similarity scores can be found in Appendix K.

4.2.4 Comparing the similarity of value summaries

When including value summaries that had their similarity scores automatically assigned, a similarity score of 4 or 5 was given in approximately 72 to 84% of cases.[15] However, since the agreement on value inclusion is high, this high agreement stems from the fact that a lot of values are agreed to not be present in transcripts. When excluding these values, results are mixed.

Table 6: Share of similarity scores assigned to value summaries, *excluding* automatically assigned scores. A and B refer to coders A and B, respectively. The models Gemma4:e4b, Phi4-reasoning, and Qwen3.6:35b are denoted with G, P, and Q, respectively. H refers to the value summaries provided by humans. Highest values marked in **bold**. Highest values compared to human judgments marked in *italic*.

Similarity score share / Coder pair	**QP**	**QG**	**QH**	**PG**	**PH**	**GH**
A % score 4-5	62	69	42	**70**	*46*	40
B % score 4-5	57	68	38	**71**	*41*	40
A % score 3-5	94	93	*71*	**96**	66	*71*
B % score 3-5	**97**	96	*74*	95	73	72

Table 6 shows the share of similarity scores assigned by coders A and B, excluding the automatically assigned scores. In roughly 38 to 46% of cases, a similarity score of 4 or 5

[15] A similarity score of 3 means the pair of summaries is roughly similar, but important details differ or are missing, 4 means the pair of summaries is mostly equivalent, but unimportant details differ, and 5 means the pair of summaries is equivalent. More details can be found in Appendix G.

is assigned when comparing the summaries provided by LLMs against those provided by humans. Both coder A and B gave Phi4-reasoning the most scores of 4 to 5, in 46% and 41% of cases, respectively. Qwen3.6 was assigned a score of 4 or 5 in 42% and 38% of cases, respectively, and Gemma4 in exactly 40% by both coders.

When including a score of 3, the results look more promising. When compared to human summaries, all models score 3 to 5 in around 70% of cases.

The similarity of provided value summaries was higher when comparing LLM pairs. Phi4-reasoning paired with Gemma4 had the highest similarity. Conversely, the lowest similarity was achieved between Qwen3.6 and Phi4-reasoning.

More detailed statistics on the assigned similarity scores are given in Appendix L.

5 Responsible Research

In terms of *human data privacy*, the only human data in this research is the dataset with transcripts between participant and an LLM, and the data extraction done by both human coders. The transcript dataset is fully anonymized and contains no personally (re-)identifiable data. Similarly, since the human coders only answered questions about the transcripts, they did not provide any personally (re-)identifiable information either.

To make sure this research is *reproducible* and done *transparently*, the choices in visualization design are justified in Section 2.1. All visualizations used are publicly available for download on GitHub.[6] To make future research using the proposed visualizations *accessible*, these files are intentionally coded in a way that they work with data extracted by the LLMs. Minimal modification of the code is needed to generate visualizations for other value models.

To eliminate bias in human coding, two coders coded transcripts independently. Only after both coders coded all conversations did they come together to discuss the results. Both the independent codes and the codes after discussing are uploaded to GitHub. To ensure the human-coded data is reliable, inter-coder reliability statistics are calculated for all dual-coded data, as described in Section 3.

Only open-source, local large language models were used. It is important to note that while this research may be used to give a participant insight into their own value model, it can also be used to create value profiles of individuals. In prior research, 85% of participants expressed concerns about companies using AI to analyze conversations and build value profiles [49]. The possibility of building value profiles is another reason only local AI models were used, and cloud models were not considered. Future work should be aware of this undesirable consequence and make sure that value profiles cannot leak to cloud providers. For example, by avoiding the use of cloud-based LLMs to extract value models.

The system prompts used for the local LLMs (found in Appendix B) and the parameters used are described in Section 3.2.5. These parameters lead to consistent outputs by the LLMs and therefore allow researchers to re-run the experiments in order to get the same results. Additionally, the LLM outputs are also provided on GitHub, so researchers do not have to re-run the local models. This makes it easier and more time-efficient to verify the results.

All calculations were performed in Microsoft Excel. The files used are available for download on GitHub as well.

LLMs may make mistakes, as shown by the justifications that had active disagreement with human justifications. Therefore, in the discussion section, the reader is warned about the consequences of over-reliance on LLMs for the task of understanding one's own value model.

5.1 Statement on the use of Generative AI

Generative AI was used as a collaboration partner in the writing of this research paper. AI was used to check the paper for spelling mistakes and to suggest improvements in writing and argumentation. However, it is important to note that AI only provided suggestions and all text has still been written by the author. In a similar fashion, AI has been used to give suggestions for improvement of the system prompt found in Appendix C. Additionally, AI helped debug LaTeX tables that were not working. It has also been used to debug the HTML files to generate the visualizations and improve the styling thereof, and to set up and debug the Python code found in Appendix D.

6 Discussion

This research aimed to find whether local, open-source LLMs ($\leq$ 35B parameters) can effectively extract value models from transcripts, to generate visualizations of these value models.

In this paper, we proposed two visualizations for value models: radar charts and value cards. Both are based on literature describing relevant properties that make visualizations effective. This literature can be found in Table 1, and the visualizations in Figures 1 and 2. We were unable to evaluate them with human participants due to delays in the ethics-approval process (see Acknowledgements). However, their theoretical backing suggests that these can aid in the reflection process, and we encourage future work to use them.
When analyzing whether LLMs can reliably extract the data needed to create these visualizations from transcripts, the main findings include:

(1) LLMs reliably identify which values are present in a transcript, closely matching human judgments ($\kappa \geq$ 0.808). This aligns with Brigola [4], extending their findings to deliberative transcripts in the public-safety domain. The LLMs are also relatively consistent (similarity score 0 in $< 10\%$ of cases) with themselves and human judgments when textually justifying a value's (un)importance. This makes them potentially useful in assisting human coders, to identify reasons why a transcript signifies the importance of certain values.

(2) LLMs are substantially less consistent with humans when ranking values, coding their importance, or summarizing their meaning ($\kappa \leq$ 0.57 for rank, $\kappa \leq$ 0.513 for importance, $\leq$ 46% equivalence on summaries). This seems to imply that LLMs struggle with more complex reasoning about the values and model as a whole. Both findings 1 and 2 are consistent with Wojtczak et al. [46], where LLMs perform well on identification tasks, but struggle on tasks requiring higher-level reasoning. We suspect these findings could be attributed in part to the *stochastic parrot* effect [3]. Specifically, we suspect the models rely on semantic correlations within their training data and match explicit keywords to flag a value's presence, without the true understanding required for ranking. Yu et al. [48] quantitatively show this effect for several LLMs, and show that further training on higher-level reasoning tasks may not significantly improve the LLMs' capabilities.

(3) LLMs are consistent in assigning ranks and importance codes to identified values. This finding reinforces the idea that finding 2 may generalize across local LLMs. However, we leave investigating this to future work.

(4) General reasoning-benchmark performance did not predict alignment with human judgment. Phi4-reasoning outperformed Qwen3.6 when compared against human judgment of ranking and importance, despite scoring substantially lower on GPQA Diamond (67.1% vs. 86%). As investigating this phenomenon is outside the scope of this

paper, we leave this open for future research. Future work could also investigate the differences between reasoning language models (RLMs) and conventional LLMs in subjective reasoning tasks, and explore more appropriate benchmarks to evaluate these capabilities.

Ultimately, these findings, which show the limitations of local LLMs in more complex reasoning tasks, highlight the need for human oversight. Using LLMs alone to extract the data required for the visualizations without human oversight, can lead to the visualizations giving wrong information. A hybrid approach, where LLMs are used for initial coding and humans perform the more complex evaluations, seems like an appropriate future direction.

6.1 Limitations

Although the relevant literature suggests the proposed visualizations could be effective, human evaluation still needs to be performed to verify this. This user evaluation could not be conducted due to ethics-approval delays (see Acknowledgements). However, a proposed method of evaluation can be found in Appendix M.

Another limitation is that the human judgments were coded by computer scientists and not by experts in human values. Future work could look at verifying the reliability of the human coding in this research by involving experts.

The dataset in this research consists of transcripts where values are often explicitly mentioned by participants. Whether LLM performance generalizes to transcripts where values are more implicit is open to investigation.

Models of up to 35 billion parameters were used in this research. Prompt tuning was performed using Qwen3.6 only. This could have improved its relative performance. Future work could investigate the use of larger models and different prompting techniques, or investigate how further fine-tuning can positively influence data extraction accuracy.

7 Conclusion

This research aimed to answer how effectively local, open-source LLMs extract information required to generate visualizations of stakeholder value models.

The first subquestion asks which properties are desirable for such visualizations. The desirable properties are provided in Table 1. Based on these, two visualizations of stakeholder value models (radar charts and value cards) are proposed, which could be effective in aiding the comprehension of a value model.

The second subquestion aims to see how consistent local LLMs are in extracting the data needed to create these visualizations, because incorrect data extraction leads to incorrect visualizations. We find that LLMs are consistent in the identification of values present in text and that they are reliable in textually justifying a value's importance, even compared to humans. LLMs are also consistent with one another in ranking these values, coding their importance, and relatively consistent in summarizing a value's meaning. However, they are not consistent with human consensus on rank, importance and summarized value meaning.

In conclusion, local LLMs are effective at identifying values present in a transcript, but their mixed performance in assigning ranks, importance, and giving value summaries shows they currently cannot reliably replace human coders. Using local LLMs to assist human coders is possible, but human oversight is still required. With the growing body of research in value-aligned LLMs, their ability to interpret and embody human values will most certainly improve. LLMs can reliably identify values in text. But, for now, the task of interpreting these human values still remains a fundamentally human endeavor.

Acknowledgements

Initially, this research was going to test the efficacy of the proposed visualizations using a questionnaire. The original plan was to do a quantitative analysis combined with a thematic analysis. Human Research Ethics approval was requested before commencing this research. However, due to administrative delays in the ethics review process, approval was not finalized within the projected timeframe. The conclusion that ethics approval would not come, came in the middle of week 6, at which point the project plan was altered to compare the data already generated by the LLMs, and to test whether LLMs give consistent justifications. This decision was made due to the limited time still available to finish the draft version of this research paper, as re-running the local LLMs takes a long time on the author's PC.

The author would like to thank Michaël Grauwde for his supervision, feedback and guidance of this research project, Willem-Paul Brinkman for his useful feedback and insights, and Sole Pera for her time and being part of the thesis committee for this research.

References

[1] Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report, 2025.

[2] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic textual similarity. In Mona Diab, Tim Baldwin, and Marco Baroni, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[4] Raoul Brigola. Vive: An llm-based approach to identifying and extracting context-specific personal values from text. Master's thesis, Utrecht University, 2024. Thesis Advisor: Davide Dell'Anna.

[5] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision to Think*. The Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann Publishers, San Francisco, CA, 1999.

[6] Studio Carreras. Values deck - studio carreras, 2025. Available online at: https://studiocarreras.com/values, Accessed: June 1, 2026.

[7] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[8] Cary Coglianese. The limits of consensus. *Environment*, 41(3):28–33, 04 1999.

[9] Fabio Crameri, Grace E. Shephard, and Philip J. Heron. Choosing suitable color palettes for accessible and accurate science figures. *Current Protocols*, 4(8):e1126, 2024.

[10] Eduardo de la Cruz Fernández, Marcelo Karanik, and Sascha Ossowski. *Value Lens: Using Large Language Models to Understand Human Values*. IOS Press, October 2025.

[11] Delft Institute of Positive Design. Design for happiness deck, 2017. Card deck.

[12] Anna K. Döring and Ariel Knafo-Noam. How do our values guide us in life? *Frontiers for Young Minds*, 7:115, 2019.

[13] Sarkpa Eastgar Garmonee and Bakare Pamela Tinashe. Conversational AI and Customer Engagement: A Comprehensive Analysis of Chatbot Technology in Modern Business Applications. *International Journal of Scientific and Management Research*, 08(10):504–515, 2025.

[14] Clement Farabet and Olivier Lacombe. Gemma 4: Byte for byte, the most capable open models. Google Blog, apr 2026. Available online at: https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/, Accessed: June 10, 2026.

[15] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Ment Health*, 4(2):e19, Jun 2017.

[16] Andrew R Gilpin. Table for conversion of kendall's tau to spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and psychological measurement*, 53(1):87–92, 1993.

[17] Arthur M Glenberg and William E Langston. Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language*, 31(2):129–151, 1992.

[18] Michaël Grauwde, Mark Neerincx, and Olya Kudina. Conversational Agents for a Deliberative Age. In *Works-in-Progress and Demonstrations track, The Eleventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP) Conference*, 2023.

[19] Irma Harms. Visualizing human values for design: Understanding, creating and analyzing human value visualizations. Master's thesis, University of Twente, Twente, NL, 2022. Available at https://purl.utwente.nl/essays/93827.

[20] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[21] Takayuki Ito, Yihan Dong, Jawad Haqbeen, Tokuro Matsuo, and Sofia Sahab. Hyperdemocracy: Towards creative consensus building between humans and ai. In *2025 IEEE International Conference on Agentic AI (ICA)*, pages 116–121, 2025.

[22] Atoosa Kasirzadeh and Iason Gabriel. In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy & Technology*, 36(2):27, April 2023.

[23] Ralph L. Keeney. Value-focused thinking: Identifying decision opportunities and creating alternatives. *European Journal of Operational Research*, 92(3):537–549, 1996.

[24] Lianne Kerlin, Michael Evans, and Phil Stenton. Digital wellbeing, 2019. Available online at: https://www.bbc.co.uk/rd/projects/digital-wellbeing, Accessed: June 1, 2026.

[25] Shadi Kheirandish, Mathias Funk, Stephan Wensveen, Maarten Verkerk, and Matthias Rauterberg. Huvalue: a tool to support design students in considering human values in their design. *International Journal of Technology and Design Education*, 30(5):1015–1041, 2019. 1015.

[26] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.

[27] Woojin Kim, Sieun Hyeon, Jusang Oh, and Jaeyoung Do. Valueflow: Toward pluralistic and steerable value-based alignment in large language models, 2026.

[28] Udo Kuckartz, Stefan Rädiker, Thomas Ebert, and Julia Schehl. *Statistik: Eine verständliche Einführung*. VS Verlag für Sozialwissenschaften, Wiesbaden, 2010.

[29] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[30] Jill H Larkin and Herbert A Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, 11(1):65–100, 1987.

[31] Netherlands Cancer Institute (NKI). Guidelines for color-blind friendly figures. Online Resource, n.d. Available online at: https://www.nki.nl/about-us/responsible-research/guidelines-color-blind-friendly-figures, Accessed: June 14, 2026.

[32] Alexandra Phillips. Colorblind safe color schemes. NCEAS Science Communication Resource Corner, UC Santa Barbara, June 2022. Available online at: https://www.nceas.ucsb.edu/sites/default/files/2022-06/Colorblind

[33] Klara Pigmans, Huib Aldewereld, Virginia Dignum, and Neelke Doorn. The Role of Value Deliberation to Improve Stakeholder Participation in Issues of Water Governance. *Water Resources Management*, 33(12):4067–4085, September 2019.

[34] Publications Office of the European Union. Accessible colour palettes. Data Visualisation Guide, 2023. Available online at: https://data.europa.eu/apps/data-visualisation-guide/accessible-colour-palettes, Accessed: June 14, 2026.

[35] Qwen Team. Qwen3.6-35b-a3b. Qwen Blog, april 2026. Available online at: https://qwen.ai/blog?id=qwen3.6-35b-a3b, Accessed: June 10, 2026.

[36] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

[37] Martin Samuelsson. Education for deliberative democracy and the aim of consensus. *Democracy Education*, 26, 04 2018.

[38] Edward R. Tufte, Dmitry Krasny, and Graphics Press. *Visual explanations: images and quantities, evidence and narrative*. Graphics Press, Cheshire, Conn., 1997.

[39] Edward Rolf Tufte. *Beautiful evidence*. Graphics Press LLC, Cheshire, CT, 2006.

[40] Barbara Tversky. Visualizing thought. *Topics in Cognitive Science*, 3(3):499–535, 2011. 499.

[41] UCL Faculty of Mathematical and Physical Sciences. Guidelines for colour blindness. PDF online publication, n.d. Available online at: https://www.ucl.ac.uk/mathematical-physical-sciences/sites/mathematical_physical_sciences/files/guidelines_for_colour_blindness.pdf, Accessed: June 14, 2026.

[42] Viktor Valgarðsson, Will Jennings, Gerry Stoker, Hannah Bunting, Daniel Devine, Lawrence McKay, and Andrew Klassen. A crisis of political trust? global trends in institutional trust from 1958 to 2019. *British Journal of Political Science*, 55:e15, 2025.

[43] Marleen van de Kerkhof. Making a difference: On the constraints of consensus building and the relevance of deliberation in stakeholder dialogues. *Policy Sciences*, 39(3):279–299, 2006.

[44] Colin Ware. *Information visualization: perception for design*. Morgan Kaufmann, San Francisco, 2nd edition, 2004.

[45] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[46] Dominika Nadia Wojtczak, Cheryl McQuire, Luisa Zuccolo, Claudia Peersman, and Ryan McConville. Performance of large language models in the cognitive analysis of misinformation: Evaluation study. *JMIR Infodemiology*, 6:e72524, May 2026.

[47] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3506–3510, New York, NY, USA, 2017. Association for Computing Machinery.

[48] Mo Yu, Lemao Liu, Junjie Wu, Tsz Ting Chung, Shunchi Zhang, Jiangnan Li, Dit-Yan Yeung, and Jie Zhou. The stochastic parrot on LLM's shoulder: A summative assessment of physical concept understanding. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11416–11431, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[49] Bhada Yun, Renn Su, and April Yi Wang. Ai and my values: User perceptions of llms' ability to extract, embody, and explain human values from casual conversations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery.

[50] Wenhao Zhu, Yuhang Xie, Guojie Song, and Xin Zhang. Eavit: Efficient and accurate human value identification from text data via llms, 2025.

[51] Ying Zhu. Measuring effective data visualization. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Nikos Paragios, Syeda-Mahmood Tanveer, Tao Ju, Zicheng Liu, Sabine Coquillart, Carolina Cruz-Neira, Torsten Müller, and Tom Malzbender, editors, *Advances in Visual Computing*, pages 652–661, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

A Transcript value model coding instructions

Coder instructions
You are annotating a set of transcripts between a host (LLM), and a participant in the domain of public safety. In each of the conversations, the LLM tries to make the participant reflect on their own values, and tries to identify values that the participant prioritizes, and what those values mean to the participant.

The value model of the LLM consists of four values:

- Autonomy: "refers to the ability of persons to create their own identity and in this way to define themselves." (Post, 2000).
- Privacy: "safeguards the spontaneous, independent, and uniquely individual aspects of the self" (Post, 2000).
- Safety: "the condition of being protected from harm (or other nondesirable outcomes) caused by non-intentional failure of technical, human or organisational factors" (van den Berg et al., 2021).
- Societal Wellbeing: refers to how an individual feels accepted or welcome in a society or community (Salehi et al., 2017).

However, you may find that some other values (e.g. transparency) are mentioned by some participants too.

Each conversation is titled RxPy where x is the round number, and y is the participant number. In case of conversations ending with part 1, part 2 the LLM briefly lost connection, so the conversation is given as two parts. Please annotate this as *one* conversation.

For each conversation, please annotate the following:

A ranking of the values mentioned in the conversation
Ask yourself: which statements does the participant give that show explicit or an implicit prioritization of values. You may find that multiple values are equally important, so you may put multiple values in the same rank.

Justifications for why the ranking was determined to be the ranking you gave
Alongside the ranking, also give justifications. In which statements does the participant explicitly say something that implicitly or explicitly prioritizes one value over another?

A label of importance for each value in the conversation: not mentioned, not important, neutral, somewhat important, very important
- *Not mentioned*: the value is not mentioned in the conversation at all, not implicitly nor explicitly
- *Not important*: the participant explicitly or implicitly shows that this value is not important to them
- *Neutral*: the participant does not state that this value is unimportant to them, but also does not state this value is very important either.

- *Somewhat important*: the participant does not explicitly or implicitly state that the value is a dealbreaker in their answer, but acknowledges that it is important to take this value into consideration when making decisions. The participant could also explicitly or implicitly state this as a secondary value to prioritize, but not as a top priority.
- *Very important*: the participant explicitly or implicitly states that this value could be a dealbreaker for making the decision at hand. (E.g. the participant explicitly or implicitly states that they feel very uncomfortable when this value is infringed upon).

Justifications for the value importance labels

For each of the values, please state all statements that lead you to ranking the value in a certain importance class.

For each value, a summary of what the value means in the context of the conversation.

Ask yourself: what does privacy mean to this person? What does the person mention about privacy in this context? Try to phrase this as bulletpoints. You may quote the participant.

And finally:

A yes/no answer whether the system came to the right conclusion about participants values in the end.

If the answer is no, please state (with justifications from the conversation) why the system was wrong.

The system may be wrong in the middle of the conversation, but if it is right in the end, the answer to this question is yes.

B System Prompt used in LLM transcript coding

Below the system prompt used to code the transcripts using LLMs Qwen3.6:35b, Phi4-reasoning:14b, and Gemma4:e4b. In all cases, this prompt was used with parameters:

- Temperature: 0
- Seed: 42
- Top_k: 1
- Top_p: 1

```
You are an expert in human values analyzing transcripts. You will be
    given a conversation between an LLM host and a human participant.

As an example, here are four common values, with definitions:

1. Safety: "the condition of being protected from harm (or other
    nondesirable
outcomes) caused by non-intentional failure of technical, human
or organisational factors"
2. Privacy: "safeguards the spontaneous, independent, and uniquely
    individual
aspects of the self"
3. Autonomy: "refers to the ability of persons to create their own
    identity
and in this way to define themselves."
4. Societal well-being: refers to how an individual feels accepted or
welcome in a society or community

CRITICAL INSTRUCTION FOR REASONING
Before generating your final JSON output, you must think step-by-step in
    plain text under a ## Step-by-Step Analysis header. Do not rush to
    conclusions. Follow this exact reasoning chain:

1. Extraction: Read the transcript and list every potential value
    expressed by the participant (including the 4 provided above or any
    new ones). Write down the exact quotes/excerpts that hint at these
    values.
2. Contextual Evaluation: For each extracted value, write a brief
    paragraph analyzing what that value actually means to this specific
    participant in the context of the transcript.
3. Weighting & Coding: Assess the absolute importance of each value on a
    scale of 1-5 (1 = very unimportant, 5 = very important). Explicitly
    write out your logic for why a specific number fits best based on the
    text evidence.
4. Comparative Ranking: Compare the values against one another. Write
    out a logical argument for why Value A is prioritized over Value B.
    Establish a valid ordinal scale (e.g., ensuring you don't skip Rank 2
    if there is a Rank 1). Multiple values can be placed in the same
    ordinal rank.

FINAL OUTPUT FORMAT:
Only after completing your step-by-step analysis above, provide your
```

```
    final conclusions structured in JSON format, like this:

{
"Privacy": {
   "Rank": int,
   "Justification_for_rank": "your justification why you ranked this
       value in this spot",
   "Importance": int (the 1-5 scale provided earlier),
   "Justification_for_importance": "your justification why you coded
       this value with this importance",
   "Value_summary": "a summary of what the value means for this
       participant"
   },

"Value_2": {...}
}

Rankings should be in a valid ordinal scale (e.g., ensuring you don't
    skip Rank 2 if there is a Rank 1).
```

C Prompt optimizing for transcript value extraction

To extract values from the transcripts, Qwen3.6:35b was used initially for its excellent reasoning qualities during testing. An initial test was performed using the following parameters:

- Temperature: 0
- Seed: 42
- Top_k: 1
- Top_p: 1

And the following system prompt:

```
You are an expert in human values analyzing transcripts. You will be
    given a conversation between an LLM host, and a human participant.

As an example, here are four common values, with definitions are:

Safety: "the condition of being protected from harm (or other
    nondesirable
outcomes) caused by non-intentional failure of technical, human
or organisational factors"
2. Privacy: "safeguards the spontaneous, independent, and uniquely
    individual
aspects of the self"
3. Autonomy: "refers to the ability of persons to create their own
    identity
and in this way to define themselves."
4. Societal well-being: refers to how an individual feels accepted or
welcome in a society or community

From the conversation given, try to extract the values of the
    participant.

Rank each value by their importance. For example, if a participant
    values privacy over safety, you would rank safety below privacy. You
    may rank multiple values in the same spot if you find one is not
    explicitly prioritized over another value.

Also give the importance of each value on a scale of 1-5, where 1 means
    very unimportant, 2 means unimportant, 3 means neutral, 4 is
    important, and 5 is very important.

Also, justify your rankings, and value importance codes by showing
    excerpts, and your conclusions from the transcript that lead you to
    this coding and ranking.

Finally, provide a summary of what this value means to the participant
    in the context of the transcript.

Structure your output in JSON format, as follows:
```

```
{
"Privacy": {
    "Rank": int
    "Justification_for_rank": "your justification why you ranked this
        value in this spot"
    "Importance": int (the 1-5 scale provided earlier)
    "Justification_for_importance": "your justification why you coded
        this value with this importance"
    "Value_summary": "a summary of what the value means for this
        participant"
    }
}
```

However, this led to the reasoning model initially thinking it should stick to the four values in the example provided. (Stating: "I should probably stick to analyzing those four, but I can include Trust if it's central, or map it to the others. I'll stick to the four provided for consistency, but I'll note Trust's role if needed.").

When coding conversation R1P1, it lead to the following (shortened) output:

```
{
  "Trust": {
    "Rank": 1,
    ...
  },
  "Privacy": {
    "Rank": 2,
    ...
  },
  "Autonomy": {
    "Rank": 2,
    ...
  },
  "Societal well-being": {
    "Rank": 3,
    ...
  },
  "Safety": {
    "Rank": 5,
    ...
  }
}
```

Here, rank 5 should be rank 4, because there should not be a rank 5 if no value is ranked 4th.

To correct for these two issues, two sentences were added:

"From the conversation given, try to extract the values of the participant. *You may find that there are also other values in the conversation. It is allowed to code these as well.*"

And

"You may rank multiple values in the same spot if you find one is not explicitly prioritized over another value. Make sure the rank is a correct scale. *You cannot have rank 2 if there is no value ranked 1.*"

Additionally, the JSON example in the system prompt was corrected to:

```
{
"Privacy": {
   "Rank": int,
   "Justification_for_rank": "your justification why you ranked this
       value in this spot",
   "Importance": int (the 1-5 scale provided earlier),
   "Justification_for_importance": "your justification why you coded
       this value with this importance",
   "Value_summary": "a summary of what the value means for this
       participant"
   },

"Value_2": {...}
}
```

This led to a correct ordinal rank, with no skipped ranks. However, Qwen3.6:35b still reasoned for adapting the value summary to the definition given, instead of giving the value in its context: "One thing: The prompt says "Societal well-being: refers to how an individual feels accepted or welcome in a society or community". I'll adapt the summary to reflect this definition in context."

To address this issue, a Chain-of-Thought reasoning pattern was added to the system prompt:

```
You are an expert in human values analyzing transcripts. You will be
    given a conversation between an LLM host and a human participant.

As an example, here are four common values, with definitions:

1. Safety: "the condition of being protected from harm (or other
    nondesirable
outcomes) caused by non-intentional failure of technical, human
or organisational factors"
2. Privacy: "safeguards the spontaneous, independent, and uniquely
    individual
aspects of the self"
3. Autonomy: "refers to the ability of persons to create their own
    identity
and in this way to define themselves."
4. Societal well-being: refers to how an individual feels accepted or
welcome in a society or community

CRITICAL INSTRUCTION FOR REASONING
Before generating your final JSON output, you must think step-by-step in
    plain text under a ## Step-by-Step Analysis header. Do not rush to
    conclusions. Follow this exact reasoning chain:

1. Extraction: Read the transcript and list every potential value
    expressed by the participant (including the 4 provided above or any
    new ones). Write down the exact quotes/excerpts that hint at these
    values.
2. Contextual Evaluation: For each extracted value, write a brief
    paragraph analyzing what that value actually means to this specific
    participant in the context of the transcript.
```

```
3. Weighting & Coding: Assess the absolute importance of each value on a
    scale of 1-5 (1 = very unimportant, 5 = very important). Explicitly
    write out your logic for why a specific number fits best based on the
    text evidence.
4. Comparative Ranking: Compare the values against one another. Write
    out a logical argument for why Value A is prioritized over Value B.
    Establish a valid ordinal scale (e.g., ensuring you don't skip Rank 2
    if there is a Rank 1). Multiple values can be placed in the same
    ordinal rank.

FINAL OUTPUT FORMAT:
Only after completing your step-by-step analysis above, provide your
    final conclusions structured in JSON format, like this:

{
"Privacy": {
   "Rank": int,
   "Justification_for_rank": "your justification why you ranked this
       value in this spot",
   "Importance": int (the 1-5 scale provided earlier),
   "Justification_for_importance": "your justification why you coded
       this value with this importance",
   "Value_summary": "a summary of what the value means for this
       participant"
   },

"Value_2": {...}
}

Rankings should be in a valid ordinal scale (e.g., ensuring you don't
    skip Rank 2 if there is a Rank 1).
```

This effectively addressed the issue, where Qwen3.6:35b reasoned: "The prompt gives 4 example values: Safety, Privacy, Autonomy, Societal well-being. I should map the extracted values to these or note new ones like Trust. I'll include Trust as a primary value since it's heavily emphasized, and map the others to the provided list or keep them as extracted." Since all issues were addressed, this is the prompt used in the research paper.

D Python code to convert transcripts from .xlsx or .csv format to .txt

The code below is used to convert the xlsx or csv files provided in the dataset at https://data.4tu.nl/private_datasets/xUIgPm6lCdMpkQef_C-AaPCswJzKgqCXdBhq2nFVIJY to a textual format. The text can then be pasted into the LLM model to extract values from.

This script is tested with Python version 3.13, openpyxl version 3.1.5, and pandas version 3.0.3.

```python
import pandas as pd
import os

INPUT_FILE_PATH = '../cnv/conversation_R1P4.xlsx'  # Update with your actual input file path
OUTPUT_FILE_PATH = '../txt/conversation_R1P4.txt'   # Desired output file path
CSV = False # Set to True if the input file is a CSV, False for Excel


def main():
    convert_excel_to_txt(INPUT_FILE_PATH, OUTPUT_FILE_PATH)


def convert_excel_to_txt(input_excel_path, output_txt_path):
    """
    Converts a conversation log Excel file into LLM-readable text file.
    """
    if not os.path.exists(input_excel_path):
        print(f"Error: Input file '{input_excel_path}' does not exist.")
        return

    print(f"Reading {input_excel_path}...")
    # Read the excel file
    try:
        if CSV:
            df = pd.read_csv(input_excel_path, sep=';')
        else:
            df = pd.read_excel(input_excel_path)
    except Exception as e:
        print(f"Error reading Excel file: {e}")
        print("Make sure 'openpyxl' is installed (pip install openpyxl).")
        return

    # Check for required columns
    required_columns = ['event', 'detail']
    if not all(col in df.columns for col in required_columns):
        print(f"Error: Excel sheet must contain columns: {required_columns}")
        print(f"Found columns: {list(df.columns)}")
        return

    output_lines = []

    for idx, row in df.iterrows():
        event = row['event']
        detail = row['detail']

        # Skip rows without message details (e.g., session_start)
        if pd.isna(detail):
            continue
```

```
49
50          # Clean HTML line breaks commonly found in conversational exports
51          detail_cleaned = str(detail).replace('<br>', '\n').replace('<br/>', '\n').
        strip()
52
53          # Format based on the sender role
54          if event == 'message_participant':
55              output_lines.append(f"### PARTICIPANT:\n{detail_cleaned}\n\n")
56          elif event == 'message_host':
57              output_lines.append(f"### LLM:\n{detail_cleaned}\n\n")
58
59      # Write to text file
60      with open(output_txt_path, 'w', encoding='utf-8') as f:
61          f.writelines(output_lines)
62
63      print(f"Successfully converted and saved to: {output_txt_path}")
64
65  if __name__ == "__main__":
66      main()
```

E Example justification similarity scores

Table 7: Example justification pairs with given codes, and reason why this code is given.

Justification pair	Code	Reason for code
(1) "The participant explicitly lists societal well-being last ('And finally Societal Wellbeing'). It is viewed as a secondary outcome or community benefit that should follow once the core triad of safety, autonomy, and privacy are secured." **(2)** "Ranks societal wellbeing last: 'The most important point out of those would probably be Safety first then actually autonomy then privacy.' "	2	Both coders describe the reason as listing or ranking societal wellbeing last, and how societal wellbeing is viewed as secondary.
(1) "Privacy is highly valued but ranked second because it is primarily framed as a protective mechanism for autonomy. The participant advocates for technical safeguards (decentralization, encryption, opt-in consent) not as an end in themselves, but to preserve individual freedom from surveillance." **(2)** "'Facial recognition in its current form infringes on privacy and should not be implemented at Woonstad Rotterdam' "	1	One justification gives the framing as a protective mechanism for autonomy, and advocation for technical safeguards. The other cites a quote that shows that privacy infringement is the dealbreaker to not implement facial recognition technology. Both justifications show importance of the value "privacy" (no disagreement on its importance), however they both give different reasons as to why "privacy" is an important value.
(1) "Valued theoretically but practically secondary. The participant claims to support autonomy and asks if it can be combined with safety, but does not advocate for it over security in this specific scenario. It is acknowledged but compromised for safety." **(2)** "After hearing the definition of autonomy participant states that autonomy is very important too, implicitly showing autonomy is valued over privacy: "i'm definitely all for autonomy", and asking if autonomy + safety is a possibility: 'can autonomy + safety be a possible combintation?' "	0	One justification argues the participant is willing to compromise on autonomy to improve safety (making autonomy a secondary value), while the other justification argues that it is a very important value, and the participant wants to combine safety and autonomy together.

F Inter-coder reliability for sentence similarity scores

Table 8: Weighted Cohen κ (rounded to 3 decimals) for justification similarity scores given by both human coders

Coder pair	**Cohen κ**
Justifications for value rank	
Including auto-coded values	
Qwen3.6:35b, Human	0.939
Qwen3.6:35b, Phi4-reasoning:14b	0.989
Qwen3.6:35b, Gemma4:e4b	0.932
Phi4-reasoning:14b, Human	0.991
Phi4-reasoning:14b, Gemma4:e4b	0.945
Gemma4:e4b, Human	0.992
Excluding auto-coded values	
Qwen3.6:35b, Human	0.879
Qwen3.6:35b, Phi4-reasoning:14b	0.980
Qwen3.6:35b, Gemma4:e4b	0.875
Phi4-reasoning:14b, Human	0.981
Phi4-reasoning:14b, Gemma4:e4b	0.904
Gemma4:e4b, Human	0.981
Justifications for value importance	
Including auto-coded values	
Qwen3.6:35b, Human	0.981
Qwen3.6:35b, Phi4-reasoning:14b	0.945
Qwen3.6:35b, Gemma4:e4b	0.956
Phi4-reasoning:14b, Human	0.977
Phi4-reasoning:14b, Gemma4:e4b	0.931
Gemma4:e4b, Human	0.992
Excluding auto-coded values	
Qwen3.6:35b, Human	0.982
Qwen3.6:35b, Phi4-reasoning:14b	0.886
Qwen3.6:35b, Gemma4:e4b	0.904
Phi4-reasoning:14b, Human	0.945
Phi4-reasoning:14b, Gemma4:e4b	0.851
Gemma4:e4b, Human	0.983

G Value Summary Similarity Coding Instructions

Coder Instructions

Two sentences may mean the same thing, even if different words and phrases are used. Conversely, two statements that are similar in their word choice, phrasing and composition may have different meanings.

You are tasked with comparing pairs of sentences and assigning a similarity score, based on their underlying meaning. (i.e. what are they saying about the participants' value)

Picture what is being described and contrast exactly what is conveyed by one statement versus what is conveyed by the other.

Do the statements refer to or describe the exact same statement, idea or thing? Or, are they similar but differ according to either large or small details?

Tips

- Be precise in the assignment of similarity scores, and try to avoid overusing any of the scores.
- Be careful of subtle differences between pairs of sentences that can have an important impact on what is being said or described.
- Ignore grammatical errors, provided value rankings or importance codes that are provided within the statements.

Please assign scores according to the examples below.

5	*The two sentences are completely equivalent, as they mean the same thing.* The bird is bathing in the sink. Birdie is washing itself in the water basin.
4	*The two sentences are mostly equivalent, but some unimportant details differ.* Two boys on a couch are playing video games. Two boys are playing a video game.
3	*The two sentences are roughly equivalent, but some important information differs/is missing.* John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.
2	*The two sentences are not equivalent, but share some details.* They flew out of the nest in groups. They flew into the nest together.
1	*The two sentences are not equivalent, but are on the same topic.* The woman is playing the violin. The young lady enjoys listening to the guitar.
0	*The two sentences are completely dissimilar.* The black dog is running through the snow. A race car driver is driving his car through the mud.

Table 9: Similarity scores with explanations and English examples from Agirre et al. [2].

H Inter-coder reliability for value summary similarity coding

Table 10 below shows the inter-coder reliability statistics for the value summary similarity scores that were assigned by human coders A and B.

Table 10: Spearman ρ, Kendall τ_b, and weighted Cohen κ (rounded to 3 decimals) for value summary similarity scores assigned by both human coders to justification pairs. A and B refer to coder A and B, respectively. Q refers to Qwen3.6:35b, P to Phi4-reasoning:14b, G to Gemma4:e4b and H to human. Highest values marked in **bold**.

Statistic / Coder pair	**A | QP**	**A | QG**	**A | QH**	**A | PG**	**A | PH**	**A | GH**
Spearman ρ						
Including values marked absent by both coders	0.986	0.987	0.989	**0.992**	0.991	0.985
Excluding values marked absent by both coders	0.700	0.753	0.762	**0.807**	0.789	0.732
Kendall τ_b						
Including values marked absent by both coders	0.955	0.954	0.954	**0.970**	0.964	0.944
Excluding values marked absent by both coders	0.663	0.692	0.682	**0.757**	0.714	0.653
Weighted Cohen κ						
Including values marked absent by both coders	0.889	0.873	0.885	**0.918**	0.898	0.863
Excluding values marked absent by both coders	0.621	0.630	0.625	**0.698**	0.634	0.613

I Cohen's κ for value inclusion

Table 11 shows Cohen κ values achieved by the different LLMs compared to human judgments.

Table 11: Cohen κ (rounded to 3 decimals) for value inclusion compared to human judgments per model. Highest values marked in **bold**.

Cohen κ / Model	**Qwen3.6:35b**	**Phi4-reasoning:14b**	**Gemma4:e4b**
Including values marked absent by both coders	0.870	0.887	**0.896**
Excluding values marked absent by both coders	0.808	0.831	**0.848**

Table 12 shows Cohen κ values achieved by different pairs of LLMs.

Table 12: Cohen κ (rounded to 3 decimals) for value inclusion in pairwise LLM comparison (Qwen3.6:35b shortened to "Qwen", Phi4-reasoning:14b shortened to "Phi4", Gemma4:e4b shortened to "Gemma4"). Highest values marked in **bold**.

Cohen κ / Model pair	**Qwen, Phi4**	**Qwen, Gemma4**	**Phi4, Gemma4**
Including values marked absent by both models	**0.948**	0.939	0.921
Excluding values marked absent by both models	**0.924**	0.914	0.887

J Detailed value rank and importance agreement statistics

Table 13 below shows the agreement statistics between value ranks assigned by LLMs compared to the triangulated human assigned ranks.

Table 13: Spearman ρ, Kendall τ_b, and weighted Cohen κ (rounded to 3 decimals) for value ranking compared to human judgments per model. Highest values marked in **bold**.

Statistic / Model	Qwen3.6:35b	Phi4-reasoning:14b	Gemma4:e4b
Spearman ρ			
Including values marked absent by both coders	0.816	**0.875**	0.853
Excluding values marked absent by both coders	0.352	**0.590**	0.392
Kendall τ_b			
Including values marked absent by both coders	0.774	**0.828**	0.792
Excluding values marked absent by both coders	0.308	**0.497**	0.307
Weighted Cohen κ			
Including values marked absent by both coders	0.841	**0.859**	0.849
Excluding values marked absent by both coders	0.536	**0.570**	0.553

Table 14 below shows the agreement statistics between value importance codes assigned by LLMs compared to the triangulated human assigned importance codes.

Table 14: Spearman ρ, Kendall τ_b, and weighted Cohen κ (rounded to 3 decimals) for value importance compared to human judgments per model. Highest values marked in **bold**.

Statistic / Model	Qwen3.6:35b	Phi4-reasoning:14b	Gemma4:e4b
Spearman ρ			
Including values marked absent by both coders	0.822	**0.855**	0.851
Excluding values marked absent by both coders	0.393	**0.486**	0.415
Kendall τ_b			
Including values marked absent by both coders	0.785	**0.814**	0.806
Excluding values marked absent by both coders	0.348	**0.421**	0.357
Weighted Cohen κ			
Including values marked absent by both coders	0.819	**0.839**	0.832
Excluding values marked absent by both coders	0.470	**0.513**	0.505

Table 15 shows the agreement statistics relating to value ranking, where the agreement between LLMs is calculated.

Table 15: Spearman ρ, Kendall τ_b, and weighted Cohen κ (rounded to 3 decimals) for value ranking in pairwise LLM comparison (Qwen3.6:35b shortened to "Qwen", Phi4-reasoning:14b shortened to "Phi4", Gemma4:e4b shortened to "Gemma4"). Highest values marked in **bold**.

Statistic / Model pair	Qwen, Phi4	Qwen, Gemma4	Phi4, Gemma4
Spearman ρ			
Including values marked absent by both models	**0.914**	0.908	0.891
Excluding values marked absent by both models	0.562	**0.583**	0.507
Kendall τ_b			
Including values marked absent by both models	**0.867**	0.861	0.832
Excluding values marked absent by both models	0.505	**0.510**	0.434
Weighted Cohen κ			
Including values marked absent by both coders	**0.909**	0.903	0.881
Excluding values marked absent by both coders	**0.704**	0.699	0.623

Table 16 shows the agreement statistics for all pairs of LLMs when comparing them in terms of the value importance codes they assigned to each value.

Table 16: Spearman ρ, Kendall τ_b, and weighted Cohen κ (rounded to 3 decimals) for value importance in pairwise LLM comparison (Qwen3.6:35b shortened to "Qwen", Phi4-reasoning:14b shortened to "Phi4", Gemma4:e4b shortened to "Gemma4"). Highest values marked in **bold**.

Statistic / Model pair	Qwen, Phi4	Qwen, Gemma4	Phi4, Gemma4
Spearman ρ			
Including values marked absent by both models	**0.922**	0.914	0.883
Excluding values marked absent by both models	**0.647**	0.657	0.515
Kendall τ_b			
Including values marked absent by both models	**0.886**	0.878	0.838
Excluding values marked absent by both models	**0.590**	0.584	0.451
Weighted Cohen κ			
Including values marked absent by both coders	**0.907**	0.893	0.869
Excluding values marked absent by both coders	**0.695**	0.664	0.591

K Similarity scores for textual rank and importance justifications

Figures 3a and 3b show the share of each similarity score assigned to human-LLM pairs for rank and importance, respectively. In over 90% of cases is a similarity score of 1 and 2 assigned. Figures 3c and 3d show the share of each similarity score assigned to each LLM-LLM pair. Here, in over 95% of cases a score of 1 or 2 is assigned.

A similarity score of 0 means that one justification argued that a value is unimportant (or low-ranked) while the other justification argues it is important. A score of 1 means that the justifications do not disagree on the importance, but they might refer to different parts of the transcript to justify this. A score of 2 means that the justifications agree on the importance and are referring to the same parts of a transcript to justify this.

	Including autocoded					
Score	A | QH	B | QH	A | PH	B | PH	A | GH	B | GH
2	76.82%	75.00%	71.82%	71.36%	70.91%	70.45%
1	15.00%	16.36%	21.82%	22.27%	21.82%	22.27%
0	8.18%	8.64%	6.36%	6.36%	7.27%	7.27%
	Excluding autocoded					
	A | QH	B | QH	A | PH	B | PH	A | GH	B | GH
2	56.41%	51.28%	38.46%	37.18%	36.71%	35.44%
1	42.31%	46.15%	61.54%	62.82%	60.76%	62.03%
0	1.28%	2.56%	0.00%	0.00%	2.53%	2.53%

(a) Share of similarity scores assigned to rank justifications

	Including autocoded					
Score	A | QH	B | QH	A | PH	B | PH	A | GH	B | GH
2	75.45%	73.18%	77.27%	75.00%	69.55%	68.18%
1	20.45%	22.73%	20.00%	22.27%	21.82%	23.18%
0	4.09%	4.09%	2.73%	2.73%	8.64%	8.64%
	Excluding autocoded					
	A | QH	B | QH	A | PH	B | PH	A | GH	B | GH
2	45.12%	39.02%	47.62%	41.67%	33.33%	29.33%
1	54.88%	60.98%	52.38%	58.33%	64.00%	68.00%
0	0.00%	0.00%	0.00%	0.00%	2.67%	2.67%

(b) Share of similarity scores assigned to importance justifications

	Including autocoded scores					
Score	A | QP	B | QP	A | QG	B | QG	A | PG	B | PG
2	77.27%	76.82%	80.00%	79.09%	80.00%	78.64%
1	18.64%	19.09%	15.91%	16.82%	15.45%	16.82%
0	4.09%	4.09%	4.09%	4.09%	4.55%	4.55%
	Excluding autocoded scores					
	A | QP	B | QP	A | QG	B | QG	A | PG	B | PG
2	48.19%	46.99%	55.95%	53.57%	56.10%	52.44%
1	49.40%	50.60%	41.67%	44.05%	41.46%	45.12%
0	2.41%	2.41%	2.38%	2.38%	2.44%	2.44%

(c) Share of similarity scores assigned to rank justifications

	Including autocoded scores					
Score	A | QP	B | QP	A | QG	B | QG	A | PG	B | PG
2	77.27%	75.00%	76.82%	75.00%	75.45%	73.18%
1	20.00%	22.27%	20.91%	22.73%	20.45%	22.73%
0	2.73%	2.73%	2.27%	2.27%	4.09%	4.09%
	Excluding autocoded scores					
	A | QP	B | QP	A | QG	B | QG	A | PG	B | PG
2	47.62%	41.67%	45.88%	41.18%	45.12%	39.02%
1	52.38%	58.33%	54.12%	58.82%	54.88%	60.98%
0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

(d) Share of similarity scores assigned to importance justifications

Figure 3: Share of similarity scores assigned to each justification pair. 2: total agreement, 1: partial, 0: active disagreement. A refers to coder A, B to coder B. The pair of justification coders is denoted with two letters. Q referring to Qwen3.6:35b, P to Phi4-reasoning:14b, G to Gemma4:e4b and H to the triangulated human justifications.

L Value Summary Similarity Scores

Figure 4 shows the number of similarity scores assigned by coders A and B, including automatically assigned similarity scores. Figure 5 shows the same, excluding the automatically assigned similarity scores. These codes are assigned according to the instructions found in Appendix G.

Code	A | QP	A | QG	A | QH	A | PG	A | PH	A | GH
0	8	11	16	17	20	13
1	1	2	8	0	5	12
2	3	4	14	3	19	11
3	25	19	22	19	14	25
4	37	36	25	32	27	22
5	146	148	135	149	135	137
% 4 & 5	0.83182	0.83636	0.72727	0.82273	**0.73636**	0.72273
% 3-5	0.94545	0.92273	0.82727	0.90909	0.8	**0.83636**
Diff 4-5, 3-	0.11364	0.08636	0.1	0.08636	*0.06364*	0.11364

Code	B | QP	B | QG	B | QH	B | PG	B | PH	B | GH
0	8	11	16	17	20	13
1	1	1	13	0	12	15
2	1	2	7	4	7	7
3	31	23	27	17	23	26
4	31	29	18	26	18	18
5	148	154	139	156	140	141
% 4 & 5	0.81364	0.83182	0.71364	0.82727	**0.71818**	0.72273
% 3-5	0.95455	0.93636	**0.83636**	0.90455	0.82273	0.84091
Diff 4-5, 3-	0.14091	0.10455	0.12273	0.07727	*0.10455*	0.11818

Figure 4: Share of similarity scores assigned by coder A and B to value summaries, *including* automatically assigned scores. The A or B in each column signifies the fact that these scores were assigned by coder A or B, respectively. Then two letters follow. Q refers to Qwen3.6:35b. P refers to Phi4-reasoning:14b. G refers to Gemma4:e4b and H to human value summaries. Below the heatmap the percentage of scores 4 and 5, 3 to 5 and the difference between the two are shown.

Code	A | QP	A | QG	A | QH	A | PG	A | PH	A | GH
0	0	0	0	0	0	0
1	1	2	8	0	5	12
2	3	4	14	3	19	11
3	25	19	22	19	14	25
4	37	36	25	32	27	22
5	10	19	7	20	6	10
% 4 & 5	0.61842	0.6875	0.42105	0.7027	**0.46479**	0.4
% 3-5	0.94737	0.925	0.71053	0.95946	0.66197	**0.7125**
Diff 4-5, 3-5	0.32895	0.2375	0.28947	0.25676	*0.19718*	0.3125

Code	B | QP	B | QG	B | QH	B | PG	B | PH	B | GH
0	0	0	0	0	0	0
1	1	1	13	0	12	15
2	1	2	7	4	7	7
3	31	23	27	17	23	26
4	31	29	18	26	18	18
5	12	25	11	27	11	14
% 4 & 5	0.56579	0.675	0.38158	0.71622	**0.40845**	0.4
% 3-5	0.97368	0.9625	**0.73684**	0.94595	0.73239	0.725
Diff 4-5, 3-5	0.40789	0.2875	0.35526	0.22973	*0.32394*	0.325

Figure 5: Share of similarity scores assigned by coder A and B to value summaries, *excluding* automatically assigned scores. The A or B in each column signifies the fact that these scores were assigned by coder A or B, respectively. Then two letters follow. Q refers to Qwen3.6:35b. P refers to Phi4-reasoning:14b. G refers to Gemma4:e4b and H to human value summaries. Below the heatmap the percentage of scores 4 and 5, 3 to 5 and the difference between the two are shown.

M Proposed method for user evaluation

Since ultimately the user-study could not be performed due to ethics approval delay, we still aim to help future research, by describing the proposed method to test these visualizations.

The idea of this user study, was to measure the comprehension of their value model arising out of a conversation with an LLM. For the conversations, the dataset by Grauwde can be used, available at https://data.4tu.nl/private_datasets/xUIgPm6lCdMpkQef_C-AaPCswJzKgqCXdBhq2nFVIJY. Transcripts from this dataset can be provided to a more user-readable chat interface look using code provided by the author on GitHub: https://maximiliaanvdv.github.io/CSE3000_LLM_Extraction_Stakeholder_Values/downloads/LLM_Data_Extraction_Supplementary_Files.zip.

The proposed method is to then ask several true-false or multiple choice questions about the transcript, and measure participants' accuracy and speed in completing these tasks. When testing for accuracy, it is also important to measure misinterpretation rates to ensure LLM errors do not mislead users.

Task workload can be measured using the NASA-TLX questionnaire [20]. The efficacy of the visualizations can be measured using the principles of accuracy, utility, and efficiency, which are described in [51].