Extracting Value, Value Tension, and Points of Agreement From Deliberation

Can LLMs identify value, value tensions, and consensus points from multi-stakeholder deliberation transcripts?

Ananya Singh[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: Ananya Singh
Final project course: CSE3000 Research Project
Thesis committee: Willem-Paul Brinkman, Michaël Grauwde, Sole Pera

Abstract

Deliberation is a process in which people come together to discuss and find solutions to complex topics. Human moderators are expensive to employ, time-consuming, and can introduce their own biases into the moderation and summary of deliberative conversations. Large language models (LLMs), on the other hand, avoid obscuring minority opinions and bring groups together rather than dividing them. In this paper, we examine whether LLMs can extract values, value tensions, and consensus points from deliberative transcripts. We experiment with 3 different prompting strategies, zero-shot, few-shot, and chain of thought prompting, and 3 different LLM models, Gemma 2, Qwen, and Mistral. We evaluate the results using ground truth annotations and an LLM as a judge study. We found that all LLMs were capable of extracting the basic constructs by providing valid outputs to the prompts given. Mistral slightly outperformed the other models according to LLM-as-a-judge, whereas Gemma 2 achieved the highest F1 score. Chain of thought prompting outperformed the others, according to the LLM-as-a-judge, with few-shot prompting achieving the highest overall F1 score. We found that the interaction between the model and prompting strategy is highly dependent on the evaluation criteria with the correlation in results between LLM-as-a-judge and metric evaluation tending to be slightly negative.

1 Introduction

Across the world, citizen assemblies are increasingly trusted to solve society's hardest questions; from Ireland's landmark abortion referendum to the conversations at the Citizens' Convention on Climate in France. Deliberative conversations are becoming the cornerstone of a fair decision-making process [27]. Despite its importance, though, meaningful deliberation is difficult to achieve in practice. A population that is diverse in its perspectives and "willing to come together are properties that do not always co-exist" [15]. As groups grow larger and more polarised, structured events become costly, difficult to scale and "can result in voices being heard unequally" [23]. Recent advances in large language models (LLMs) have opened a new possibility in the context of deliberation. Tessler et al. [23] have shown that AI mediators leave groups comparatively less divided, produce more agreeable statements and avoid discriminating against minority perspectives compared to their human counterparts.

Previously, LLMs have been explored as thinking partners alongside humans [8], as thinking assistants for self-reflection [19], and as AI supporting stakeholders in surfacing and improving fairer decision-making processes [29]. Babatunde et al. [2] also explored extracting arguments and categories from deliberation using various prompting techniques. This work classifies individual utterances into fixed categories, and we aim to build upon this to extract open-ended constructs that require reasoning across multiple turns. Our research, therefore, focuses on the systematic extraction of values, value tensions, and consensus points as distinct constructs from deliberative transcripts. This gap in research is important to fill since normal summaries risk obscuring the underlying disagreements that matter and degrading minority opinions [10]. Additionally, deliberative forums should look not only at points of agreement in a conversation but also at the "tensions between competing values and conflicting priorities" [5]. Bridging this knowledge and exploring the use of LLMs in this context can consequently help make decision-making processes fairer and make summarisation and moderation more cost-effective and efficient. As a result, this research paper answers the broad research question: *Can large language models identify value, value tensions, and consensus points from multi-stakeholder deliberation transcripts?*

The sub-questions derived are the following:

- *SQ1: Which LLM most accurately extracts deliberative constructs from multi-stakeholder transcripts, as measured by metric evaluation and an LLM-as-a-judge evaluation?*
- *SQ2: Which prompting strategy most reliably elicits correct and complete deliberative construct extraction, as measured by metric evaluation and an LLM-as-a-judge evaluation?*
- *SQ3: How does the interaction between model and prompting strategy affect extraction quality?*

As a result of these questions, the contributions of this paper are: (1) a systematic comparison of 3 LLMs and 3 prompting strategies for deliberative construct extraction, (2) a LLM-as-a-judge study with results across four quality criteria, (3) a metric evaluation comparing the precision, recall and F1 scores derived from ground truth annotations, (4) a benchmark of model-prompt combinations ranked by extraction quality for deliberative content.

The rest of the paper is structured as follows: we walk through related work in deliberative theory, the use of LLMs as assistants, and their direct use in deliberation (Section 2). Then, we motivate the design choices and structure of the methodology (Section 3), followed by the methodology's execution (Section 4) and the outcomes and results (Section 5). We then discuss the implications of this research for ethics and reproducibility (Section 6). We go over the results in more depth and what they mean for the broader field of deliberation. Finally, we conclude the paper with an overview of its contents, the results of the questions and future work (Section 8).

2 Related Work

2.1 Deliberative Theory

In this research, we follow Chambers's [6] definition of deliberation, defined as a "debate and discussion aimed at producing reasonable, well-informed opinions in which participants are willing to revise preferences in light of discussion, new information, and claims made by fellow participants. Although consensus need not be the ultimate aim of deliberation, and participants are expected to pursue their interests."

This definition alone, however, is insufficient as a basis for extracting the constructs we defined. Carcasson and Sprain [5] argue that deliberation should look beyond seeking agreement but also highlight the tensions between opposing and adverse values and opinions. This motivates our focus on extracting not only the consensus points but also values and value tensions as distinct constructs to capture the full deliberative exchange rather than just conclusions.

As for our constructs, we adopt Schwartz's [21] definition of values as "a trans-situational goal that serves as guiding principles", thus defining a value as a principle that guides a person's judgment about what is important. In the context of deliberation, these values motivate the positions participants take and serve as justifications [21]. Identifying values allows us to analyse conflicts and agreements at a fundamental level, as they form the basis of stakeholder intent. Following the definition of a value, a value tension is a conflict between two values that are each independently legitimate; tensions can surface when participants invoke different values in response to the same issue. Finally, a consensus point is a claim that

receives endorsement from all (or a subset) of participants in a deliberative context. This, however, does not require unanimity but instead overlapping agreement across otherwise divergent positions [20]. Consensus points form the baseline summary of the deliberation outcome; these points highlight shared ground and provide a foundation for collaborative decision-making.

2.2 LLMs as thinking partners/assistants

Before applying LLMs specifically to deliberative contexts, it is worth noting that LLMs have already demonstrated the capacity to engage meaningfully with open-ended reasoning tasks in collaboration with humans. Collins et al. [8] have advocated for a "design that explicitly recognises and engages with the richness and diversity of human thought in an often unpredictable world". They investigated the context for using LLMs as "thought partners" (collaboratively ideating and thinking together) in programming, embodied assistance, storytelling, and medicine. Their work shows that LLMs can meaningfully participate in tasks that require an understanding of the nuances of different human opinions. To extend this further, Park et al. [19] have explored LLMs as conversational assistants, which they call "thinking assistants". In their work, they designed an agent that would ask "thought-provoking" questions to enhance the user's ability to engage in possibly difficult decision-making processes. They concluded that participants think reflectively in the presence of the conversational agent more than they do without it. This matters for our work because it suggests that LLMs are not merely tools for summarising deliberative conversations but have a real grasp of the values and opinions that humans can hold and express.

These findings are encouraging and suggest that LLMs can reason with humans about values and reflective work, but they carry an important distinction. Both studies involve working with LLMs alongside humans in real time. Our task, however, is different since the LLM must reason over a completed transcript without any support or interaction from the speakers. It is therefore not guaranteed that the capacity for collaborative reasoning demonstrated in these studies transfers to the task we are trying to execute. However, they do establish that LLMs understand human values and intent, which is a necessary precondition for our work to proceed.

2.3 LLMs and agents in deliberation

Building on this capacity for reasoning, LLMs and agents have also been explored more directly in the context of deliberation. Agarwal et al. [1] have explored the use of conversational agents to enforce "accurate belief formation", particularly on WhatsApp groups. They explored the design and implementation of an agent to help users in WhatsApp groups deliberate about harmful content, particularly misinformation and disinformation. They found that participants found an external agent that would initiate conversations about harmful content useful. Additionally, Babatunde et al. [2] investigated how LLMs can moderate large-scale online deliberations as a scalable and inexpensive alternative to human moderators. LLMs were prompted to extract answers and classify them into predefined categories such as metrics, barriers, and arguments. They used zero-shot prompting, where the model is provided with a description of the task only, 2 different few-shot techniques, where the model is provided with the task annotated examples, as well as the description and few-shot chain of thought prompting, where the model is provided with annotated examples, as well as reasoning steps to follow. They found that few-shot prompting improved the model's

ability to identify and categorise answers, but not arguments. Whereas chain-of-thought prompting increased complexity, resulting in negatively impacted results.

3 Research Motivation and Approach

3.1 Choosing a dataset

Several existing datasets were considered, including Habermas Machine, Europolis, DeliData, and MeetingBank; however, none fully met the definition of deliberation defined in 2.1. These datasets were either not in the right context, such as DeliData [16], contained only summaries of the deliberation or included a lot of buffer speech. As a result, we make use of the UK House of Commons (Hansard) dataset for prompt tuning. This dataset was chosen because it is multi-stakeholder, with a diverse set of participants representing different interests and positions. Debates are structured around specific policy topics, and participants are expected to justify their positions. Finally, as a public record, Hansard ensured full reproducibility of our work.

When compiling all the different Hansard transcripts, four themes emerged: terrorism, surveillance, crime and policing, and emergency services. These topics contained many sub-topics, of which we filtered out all topics that were too short or not in the correct context (more procedural than actual deliberation). Amongst the remaining transcripts, we chose 5 (one from each theme, plus an extra one from crime policing) across distinct policy topics, including online safety, medicine, work for prisoners, terrorism, and data privacy. This ensures that results are not specific to a single policy domain and allows for a more generalisable assessment of extraction quality. For the purposes of evaluation, we artificially created 6 deliberative transcripts intended to represent a real deliberative conversation among stakeholders in the context of public safety. We decided to create our own transcripts for the 'test' set so we could tailor the content to the specific realm of public safety. These transcripts were then checked by 2 annotators to ensure that personal biases were not included in the content and that the artificial transcripts accurately represent a normal conversation. To avoid introducing personal bias into the evaluation pipeline, the training and test sets are deliberately kept separate. This separation ensures that any assumptions that shaped our prompt tuning can't leak into what we reward on the test set.

3.2 Prompting Strategies

Sclar et al. [22] found that the performance of a large language model is highly sensitive to the formulation of prompts. They found that "performance spread caused by arbitrary prompt formatting choices may influence conclusions made about model performance", suggesting that no single prompting strategy can conclusively show a model's capabilities. Language model performance varies across tasks, metrics and even evaluation methods; no single benchmark can accurately map the reliability of the LLM [17]. Different prompting paradigms will, therefore, provide the model with varying levels of guidance and allow us to isolate how much contextual guidance is needed for the model to extract deliberative constructs.

For this reason, it is important to capture the full extent of a model's capabilities. We are first using zero-shot prompting, which requires prompting an LLM with a description of the task without examples. This strategy is particularly useful for evaluating an LLM's performance, since its output depends solely on the model's pre-training [13]. While this

strategy is a good benchmark to compare models, we are providing a specific task to the LLM which requires a high level of reasoning; for this reason, we also introduce few-shot learning, which aims to ground the LLM in the task by providing a description as well as annotated examples, and can help performance by increasing the amount of context [4]. Finally, we are using chain of thought (CoT) prompting, which involves prompting the LLM to think in steps, eliciting a more structured and thoughtful response, appropriate for multi-step reasoning across speakers, since chain of thought prompting is useful for reasoning in general [26].

We also considered two other possible techniques: step-back prompting and self-consistency prompting. Step-back prompting is when the LLM is prompted to 'step back' and take a higher level abstraction of the principles of the task before answering the prompt [30]. This may seem appealing at first, since it would allow for deeper reasoning, but stepping back to a higher level of reasoning can cause the LLM to lose details from the provided transcript. Secondly, we also considered self-consistency prompting, which "enhances the reasoning capabilities of LLM by generating multiple reasoning paths and selecting the most consistent answer". However, we did not continue with this approach because it can be less effective for open-ended tasks, making it unsuitable for this task [3].

3.3 Model Selection

While we are using different prompting strategies to evaluate a single model's capabilities, the generalisability of patterns observed in extraction quality is yet to be explored. In a field where model performance is changing rapidly and a model considered cutting-edge when it was released can be outdated just months later, it is important to assess whether findings about deliberative construct extraction are likely to hold for future models as well.

For this reason, we selected 3 models, allowing us to assess whether extraction patterns are model-specific or reflect broader trends across LLMs. For a model appropriate for this type of research, we would want a general-purpose or reasoning model that works well with open-ended tasks. We are using Ollama for open-weight models for reproducibility. We first chose Qwen 2.5 7B, which is considered a stronger general-purpose instruction-tuned baseline. We are also working with Mistral 7B, at a scale similar to Qwen, and it is often used as a benchmark model in prompting research [14]. Finally, we are using Gemma 2 9B, which is slightly larger than the other two and has strong instruction-following capabilities. We decided to explicitly choose LLM from 3 different providers (Alibaba, Mistral and Google).

3.4 Problem and Method

Given a multi-stakeholder deliberative transcript, the task is to identify and extract the 3 constructs defined above. Unlike surface-level summarisation, this task requires mapping out the underlying principles that motivate each speaker's position, detecting conflicts, and identifying moments of agreement for future decision-making. This makes it a challenging open-ended reasoning task with no single correct answer.

We first start by annotating our training dataset of deliberative transcripts with value, value tensions, and consensus points. These will serve as our ground truths for prompt tuning. Next, we tune our prompts using the ground truths. These finalised prompts are used on the selected models on a test set of deliberative transcripts. To evaluate the quality of the extractions, we again compare them to ground truth values, as well as conduct an LLM-as-a-judge study on a preset criterion.

4 Method

4.1 Annotation

In order to ensure that the prompts that we are using are properly tweaked to the task at hand, we will be making use of a train set of transcripts, validating on a test set of transcripts. For this, we will need to establish the ground truth annotations for the transcripts so that we can compare the LLM outputs to them. Simply having a single researcher annotate all points in the dataset can introduce annotation bias, which arises from an annotator's predisposition to label and interpret data differently from each other [9]. To prevent this, we had 2 coders labelling the dataset separately to prevent influencing each other's process. After annotating, the coders met in person to reach a consensus on the final ground truth annotations and discuss any disagreements they had about specific annotations. The coders were provided with instructions, which can be found in Appendix A. Using normal metrics for inter-annotator agreement (IAA) such as Cohen's κ or Krippendorf's α for this annotation task would not be a suitable metric since this annotation task is open-ended and does not have a fixed amount of labels or utterances, which in turn can make Cohen's kappa unreliable as the number of unlabeled utterances increases [12]. For this reason, we decided to use a span-based agreement metric adapted from the work of Hripcsak et al. [12], where they argue an "F-score" approach is more appropriate for reasoning tasks. As a result, we compute the pair-wise agreement between the 2 annotators' coding on the constructs and labels given. Additionally, we calculated the survival rate (how many of the constructs actually ended up in the final ground truth labels) after reaching consensus on the ground truth annotations.

-	Both Coders	Coder 1 only	Coder 2 only
Constructs Found	18	14	20
Included after Triangulation (%)	94.4	64.3	45.0

Table 1: Distribution of constructs identified for all 5 training transcripts and percentage of constructs that were included in triangulation

-	Both Coders	Coder 1 only	Coder 2 only
Constructs Found	28	43	13
Included after Triangulation (%)	100.0	86.0	23.1

Table 2: Distribution of constructs identified for all 6 test transcripts and percentage of constructs that were included in triangulation

From Table 1, we can use the adapted F1 score metric from Hripcsak et al., where we calculate the score from the perspective of Coder 1 using Coder 2 as a reference to get an F1 score of 0.514. We can also see that after both annotators came together to triangulate results for the final ground truth annotations, most constructs coded by both coders were included, as well as a portion of independently coded constructs. Looking at the test transcript reliability, we can see that it is slightly lower at 0.5, although still similar.

4.2 Prompt Tuning

Prompt tuning was conducted over 3 iterations using the Hansard dataset. Each iteration followed the same cycle: run all 3 prompting strategies across all transcripts, compute pre-

cision, recall and F1 against the ground truth annotations, which is a standard approach to compare model outputs [25], identify failure patterns and revise the prompting accordingly. We decided to prompt tune on the Llama 3.1 8B model, intentionally kept separate from the 3 models we are testing to ensure that the prompts aren't specifically designed to favour a specific model. The tuning process surfaced 3 persistent challenges: the models consistently misclassifying value tensions as points of internal deliberation (e.g a person weighing two options amongst themselves); consensus points frequently identified without multi-speaker endorsement; and many values identified having no basis. Tuning was concluded after iteration three, when marginal F1 gains diminished, and remaining weaknesses could be attributed to model architecture. The prompts we started with can be found in B.1, and the criteria that we use to evaluate the correctness of the outputs to the ground truth labels can be found in Appendix C.1.

Iteration One: In this iteration, the most glaring problem we noticed was that out of the 5 transcripts, CoT prompting only yielded two successful outputs, and a few shots completed the task successfully only 3 times. Table 3 and Table 4 contain the per prompt and per construct precision, recall and F1 scores. Value tensions have a low precision, meaning we should reinforce the prompt when it comes to its definition. Constructs also have a low F1 score, which can be attributed to the fact that many consensus points identified correct points, but did not have multiple speakers involved in their justification.

	Precision	Recall	F1 Score
Zero Shot	0.353	**0.295**	0.321
Few Shot	**0.480**	0.279	**0.353**
CoT	0.462	0.200	0.279

Table 3: Iteration 1 Prompting Metrics

	Precision	Recall	F1 Score
Value	**0.744**	0.312	**0.440**
Value Tension	0.111	**0.429**	0.176
Consensus Point	0.174	0.118	0.141

Table 4: Iteration 1 Construct Metrics

Following from this analysis, we made changes to reinforce the definitions of the constructs and the desired outputs; full changes can be found in Appendix C.2.

Iteration Two: In this iteration, all 5 transcripts for all prompting strategies were successful. Consensus point extraction metric improved; however, value recall dropped, which could indicate that the tightened definitions improved correctness at the cost of coverage, as seen in Table 6 and Table 5. From this analysis, we should add more explicit guidance for detecting values, provide more nuanced examples for few-shot settings, and introduce examples to CoT with full changes specified in Appendix C.3.

Iteration Three: All prompting strategies continued to produce valid outputs, with no prompting strategies failing to complete the task. From Table 8 and Table 7, we recovered the behaviour of CoT prompting from its decline in iteration 2. Though value tensions remain the weakest construct, fewer false positives were noticed in this iteration.

	Precision	Recall	F1 Score
Zero Shot	**0.581**	**0.290**	**0.387**
Few Shot	0.545	0.279	0.369
CoT	0.417	0.200	0.270

Table 5: Iteration 2 Prompting Metrics

	Precision	Recall	F1 Score
Value	**0.826**	0.204	0.327
Value Tension	0.167	0.375	0.231
Consensus Point	0.542	**0.382**	**0.448**

Table 6: Iteration 2 Construct Metrics

	Precision	Recall	F1 Score
Zero Shot	0.540	0.317	0.400
Few Shot	0.583	**0.333**	**0.424**
CoT	**0.636**	0.233	0.341

Table 7: Iteration 3 Prompting Metrics

	Precision	Recall	F1 Score
Value	**0.906**	**0.309**	**0.460**
Value Tension	0.125	0.230	0.167
Consensus Point	0.417	0.303	0.351

Table 8: Iteration 3 Construct Metrics

Based on these observations, prompt tuning was concluded. It should be noted that, for zero-shot prompting, we clarified the expected structure for value, value tensions, and consensus points (without providing examples to adhere to the zero-shot prompting structure).

4.3 Creating the test transcripts

The decision to construct rather than source the test transcripts was motivated by two main points. First of all, there was no publicly available dataset that fulfilled the criteria set in Section 2.1 and was set in the public safety domain. Secondly, using an artificially created set strictly for the test set will ensure that the same assumptions that shape the training set annotations do not implicitly shape what the test set rewards. A risk of creating synthetic transcripts, though, was researcher bias in content construction, which can result in transcripts that are too neatly structured and may not generalise to real deliberative conversations. To combat this, the transcripts were created using real research done on the topics covered in the transcripts. Transcripts were created with varying amounts of constructs; some conversations have value, value tensions, and consensus points present, whereas others seemed to reach no agreement at all. Additionally, to ensure content validity, a 2-person review process was conducted before the use of transcripts. 2 reviewers, independent of the process of creating the test transcripts, were provided access to the transcripts and asked to identify any moments that felt unnatural, implausible, or inconsistent with genuine multi-stakeholder deliberations and possible changes to make. These reviewers independently assessed the transcripts, and passages flagged by both reviewers were revised. The instructions for the reviewers can be found in Appendix D.

4.4 Running the LLMs on the transcripts

Now that we have the final test transcripts, with the final prompts and the models selected, we ran every combination of model, transcript and prompt together, yielding a total of 54 outputs, 9 outputs per transcript.

To ensure that we had the most deterministic and 'straight to the point' answers possible, we set the temperature to 0, which lets the LLM give more consistent and comprehensible responses. Additionally, we set the top_p (nucleus sampling) to 1.0, which ensures that the pool of words considered by the LLM in the response is the top choice probability-wise. Finally, we set the top_k as 1 to ensure that the model always picks the most likely word.

4.5 Measures

In Section 4.1, we described the processing of obtaining ground truth labels for a metric evaluation, from which we will calculate precision, recall and F1 scores. In addition to an evaluation with ground truth tables, we will also conduct a LLM-as-a-judge study. Liang et al. [17] have shown that a single evaluation metric does not capture the full performance of

the LLM, so we will use LLM-as-a-judge to evaluate the quality of the content of the output. LLM as a judge is a process where a usually larger LLM is used to evaluate the responses of our smaller LLMs. LLM rankings have been shown to often agree with human rankings and can evaluate each output independently of each other [7].

Therefore, we decided to use the Deepseek R1 14B model since it was a larger model (with 5-7 billion more parameters than the other models we used), had a different provider than the models we used before, and is known to be a reasoning model. Defining an evaluation criterion is imperative for LLM-as-a-judge [11], thus we provided the LLM definitions for the 4 criteria we were evaluating on: faithfulness, construct correctness, coverage and label quality. The full prompt can be found in Appendix E. We chose these criteria specifically to ensure that all output responses were not hallucinated, capture at least the main points created in the conversation and fit the constructs defined. These criteria were to be rated on a Likert scale of 1 - 5. Additionally, LLM as a judge can contain many risks such as positional, verbosity and self-enhancement risk [31]. We addressed positional bias, where LLMs can prefer responses based on when they see the response, by randomising the position of the outputs before letting the LLM evaluate them. We did not believe the verbosity risk, which is when LLMs prefer larger responses, to be within the scope of this experiment since all responses were of a similar size and followed a strict format. Finally, for self-enhancement risk, where LLMs prefer responses created by them, we used a completely different LLM (Deepseek 14b) to evaluate rather than to create responses, which mitigates this risk. Finally, we ran LLM as a judge using 10 different seeds in order to decrease the amount of randomness for the evaluations.

5 Results

5.1 Effect of Model

We first computed the metric evaluation for each model across all prompting techniques and all 6 transcripts, shown in Table 9. For the LLM-as-a-judge, we took the mean of all the Likert scores (for each criterion and overall) across all 10 seeds, shown in Table 10.

Model	Precision	Recall	F1
Gemma	**0.596**	**0.492**	**0.540**
Qwen	0.555	0.365	0.441
Mistral	0.511	0.449	0.478

Table 9: Per Model scores for metric evaluation

Model	Faith.	Corr.	Cover.	Label Q.	Overall
Gemma	4.467	**4.739**	4.322	4.733	4.615
Qwen	4.644	4.678	4.350	**4.700**	4.593
Mistral	**4.672**	4.700	4.417	4.689	**4.619**

Table 10: Per Model scores for LLM-as-a-judge

Model	Gemma	Qwen	Mistral
Metric Variance	0.106	0.138	0.101
LLM-as-a-judge Variance	0.039	0.027	0.018

Table 11: Variance in scores for metric and LLM-as-a-judge per model

As well as analysing the raw scores for both the LLM-as-a-judge and metric evaluation, we also observed the variation in scores. We calculated the variance across all 6 transcripts for the F1 scores, and the variance against the average Likert scores across all 10 seeds for

all models, as seen in Table 11. Overall, we can see that Gemma performs slightly better than the other models as per metric evaluation, followed by Mistral and finally Qwen. LLM-as-a-judge, however, slightly disagrees with the ranking, stating that Mistral was the model that performed the best overall, followed by Gemma, with Qwen working the worst out of the 3. We can also see that the variance in F1 scores is around 0.1 with a sample size of 6 transcripts, indicating moderate noise. On the other hand, LLM-as-a-judge has very low variance between scores. Additionally, looking at the final F1 scores per model, Gemma has the highest F1 score in 3 of the 6 transcripts, Qwen in 2, and Mistral in 1.

5.2 Effect of Prompting Strategy

Prompt	Precision	Recall	F1
Zero-shot	0.565	0.565	0.565
Few-shot	0.523	**0.732**	**0.610**
CoT	**0.581**	0.429	0.493

Table 12: Per Prompt scores for metric evaluation

Model	Faith.	Corr.	Cover.	Label Q.	Overall
Zero-shot	4.628	**4.717**	4.322	4.706	4.593
Few-shot	4.561	4.694	4.322	4.656	4.558
CoT	**4.794**	4.706	**4.444**	**4.761**	**4.676**

Table 13: Per Prompt scores for LLM-as-a-judge

Model	Zero-shot	Few-shot	CoT
Metric Variance	0.124	0.082	0.073
LLM-as-a-judge Variance	0.034	0.028	0.025

Table 14: Variance in scores for metric and LLM-as-a-judge per prompt

We can see in Table 12 that metric evaluation shows that few-shot prompting performed the best, followed by zero-shot and finally CoT. Interestingly, though, LLM-as-a-judge disagrees with this ranking, placing CoT first, followed by zero-shot (in agreement) and finally few-shot, which are completely opposite results from each other. In Table 14, we can also see that the variance for LLM-as-a-judge has slightly increased overall, whereas for metric evaluation, the variance has overall decreased. Finally, looking at the F1 scores transcript-wise, few-shot prompting achieved the highest F1 score in 3 out of the 6 transcripts, with zero-shot prompting achieving the highest score in the remaining 3 transcripts.

5.3 All Model Combinations

Combination	Gemma Zero-shot	Mistral Few-shot	Gemma Few-shot	Gemma CoT	Mistral CoT	Qwen Zero-shot	Mistral Zero-shot	Qwen Few-shot	Qwen CoT
Overall Score	0.576	0.526	0.525	0.521	0.455	0.454	0.451	0.427	0.396

Table 15: Overall F1 scores for each model-prompt combination for metric evaluation

Combination	Gemma CoT	Qwen CoT	Mistral CoT	Mistral Zero-shot	Gemma Zero-shot	Qwen Few-shot	Mistral Few-shot	Gemma Few-shot	Qwen Zero-shot
Overall Score	4.700	4.671	4.658	4.638	4.625	4.592	4.562	4.521	4.517

Table 16: Overall Likert scores for each model-prompt combination for LLM-as-a-judge

We also explored if there was any correlation between the LLM-as-a-judge and metric rankings as seen in Table 15 and Table 16. We decided to calculate Kendall's τ which is suitable to measure the ordinal association between rankings, particularly for small sample sizes since we only have 9 data points. The Kendall's τ is -0.11 which indicates slight negative correlation.

6 Responsible Research

6.1 Reproducibility

All code, annotation data, transcripts and LLM outputs can be found in the following repository [1]. The training dataset used was deliberately chosen to be publicly available, with the selected transcripts present in the database. All LLMs chosen are also open source and openly runnable with all parameters present in the provided code as well as in Appendix F. All LLM outputs, as well as data for the two coders and calculations for precision, recall and F1 score, are presented in an Excel file in the database. Finally, if we look at the reproducibility in terms of the inter-annotator agreement, we can see a similar agreement rate as seen by the F-score for the training and test transcript annotations. Since the coding task was an open-coding task with no fixed labels, it is logical to assume more disagreement and false negative labels.

6.2 Societal Impact and Ethics

Regarding the ethics of this research, human data were used to annotate the test and training sets with constructs. In this case, no personal information of the annotator was provided or stored. Additionally, since we are not experts in deliberative AI, the transcript annotations and interpretations of LLM outputs should be viewed with caution and may be limited by our understanding of the field. The open source and open weight models were also chosen for the reason of data security, since locally run LLMs on Ollama do not send data to the cloud or third-party services, which adds a blanket of security as opposed to using cloud-based models such as GPT. Concerning the use of AI, generative AI tools were used to enhance the clarity and readability of written text. Prompts used were written to improve suggestions to the written text and not the synthesis of new text. All research activities, including literature review, methodology design and analysis, were conducted by the author. AI-generated suggestions were critically reviewed and modified where necessary.

Deploying this in a real pipeline raises important questions. The models tested here are small and run on synthetic transcripts. To adapt this work responsibly, further work on real citizen assemblies should be continued to better adapt to real scenarios. Algorithmic bias is also a concern; if a model consistently underrepresents certain opinions, it can quietly skew which points are brought up to question. Thus, real deployment of LLMs in this space

[1] `https://github.com/ananya290905/RP_Project`

should be done with human oversight. This is further supported when we look at the inter-reliability between the annotators as seen by the F-score compared to the F1 score of the LLMs. The F-score of the inter-annotator reliability was 0.5 and 0.514 for the test and train transcripts, respectively, as compared to the LLM outputs' F1 scores; the highest achieved score is 0.576, only slightly higher than humans, which indicates that this extraction task cannot be fully offloaded to LLMs.

7 Discussion

7.1 LLMs for extracting constructs

LLMs have long been explored as a scalable alternative to human moderators and summaries, for their efficiency and comparative lack of bias against minority opinions. The results of this study suggest that LLMs can meaningfully extract deliberative constructs from multi-stakeholder transcripts.

Across both evaluations, all 3 models were capable of extracting the basic value, value tensions, and consensus points, with Gemma 2 achieving the strongest performance on the F1 metric (0.540) and Mistral receiving the highest overall Likert score (4.620). An explanation for Gemma 2 getting a slightly higher F1 score than the other models could seem to be attributed to model size. Zadenoori et al. [28] found that models with over a trillion parameters only work slightly better than smaller models, which suggests that the parameter sizes do not have an effect on the results. Although from Table 11, we can see that the variance observed across transcripts is around 0.1 to 0.14, so this difference relative to Qwen (F1: 0.441) may not be substantial. While Gemma 2 achieved the highest F1 score aggregated and was the top performer in 3 out of 6 transcripts, this does not represent a dominant or consistent advantage. Qwen performed the highest in 2 transcripts despite having the lowest aggregated F1 score, indicating that performance was more variable across transcripts. Mistral, despite only performing the highest in 1 transcript, maintained the second-highest F1 score overall, indicating more consistent performance. As for the LLM-as-a-judge, we see little variance in the Likert scores, with all models ranking an average above 4 out of 5. This may be attributed to leniency bias, which is where the LLM acting as the judge is overly lenient when it comes to evaluating the responses and defaults to giving high scores to all responses [24]. Observing F1 scores, while there isn't a specific chart to judge how well a score is, Murton et al. [18] found that in their task of extracting data from randomised clinical trials, LLMs achieved an F1 score of 0.85, which in their case was "often equaling human performance". In our case, the F1 scores for all combinations ranged from 0.576 to 0.396, which is considerably lower, and could be attributed to the higher reasoning level required to extract deliberative constructs rather than medical data. However, looking at that source, we can say with some certainty that the accuracy of extraction was not on par with that of human performance. Additionally, if we compare our F1 score to that of Babatunde et al.'s [2], they found that few-shot achieved the highest F1 scores (when looking at the better few-shot prompt), followed by zero-shot and finally CoT prompting in mining tasks. This closely mirrors our results when comparing F1 scores of the different prompting strategies, where we received a similar ranking.

When it comes to the different prompt techniques, however, we can see that for the metric evaluation, few-shot prompting works the best, although for LLM-as-a-judge, CoT prompting achieved the highest average Likert score. What is interesting, though, is that although CoT prompting achieved the highest scores according to LLM-as-a-judge, it received

the lowest F1 score, that too by a substantial amount. Wei et al. [26] had discovered that CoT prompting works best with larger (100B +) models, which the models we used are not. Additionally, the results we found in metric evaluation align closely with Babatunde et al. [2] work in extracting arguments and argument classification, where they found that few-shot learning achieved the highest performance for their extraction tasks, whereas few-shot CoT prompting added complexity to the tasks, which negatively affected the model's ability. This is highly relevant to our results as well, since they also worked with precision and F1 scores and could explain why CoT prompting performed the worst in metric evaluation. CoT prompting is still valuable for reasoning and logical deduction, though, which could explain why it scored high on the Likert scale. The outputs from CoT were still valuable, and the phrasing and content of the outputs (possibly fluff) yielded high scores.

Although we looked at prompting strategies and models separately, we also analysed the final rankings from the metric and LLM-as-a-judge evaluation, where Gemma 2 with zero-shot prompting achieved the highest F1 score, and Gemma 2 with CoT prompting received the highest average Likert score. When looking at Kendall's τ, though, we can see a slight negative correlation between the two rankings. This may be attributed to the fact of the LLM judge's leniency, but also could be due to the fact that both methods evaluate on different criteria; the metric evaluation focuses completely on the coverage and accuracy of the construct extractions, whereas LLM-as-a-judge takes a more subjective measure by evaluating on the quality and actual content of the outputs.

7.2 Limitations

This study provides the groundwork and intuition for using LLMs in the deliberative space for the extraction of value, value tensions, and consensus points, but several limitations of the current scope of the work suggest gaps for future work. First, the use of synthetic transcripts rather than real deliberative conversations could have impaired the quality of the extraction by still being too neatly structured compared to real discussions. This can limit the generalisability of the results to real-world deliberative conversations. We also limited the scope of this experiment to 3 prompts and models, which provides a gap for future work to explore further techniques and more model architectures to make a more conclusive and generalisable decision on the best model and prompt strategy to be used for this task. Finally, with the use of LLM-as-a-judge as an evaluation method, we saw possible leniency bias when it came to evaluating every output. This could possibly mean that the ratings of an LLM may not align with those of a human evaluation study. This can already be seen when looking at the correlation between metric and LLM scores, where there is a slight negative correlation. Finally, as mentioned in the responsible research section, we are not experts in the field of deliberative AI, meaning that interpretation of outcomes and annotations of ground truth labels may contain biases or inaccuracies and should therefore be interpreted with appropriate caution.

7.3 Implications for deliberative contexts

Beyond technical findings, this study has broader implications for the use of LLMs in real deliberative contexts. As deliberative conversations and citizen assemblies scale in size, the cost and time consumption of human moderation becomes a bottleneck to overcome. The results in this study suggest that LLMs can serve as a useful tool to assist and extract basic constructs from a transcript in an effort to not replace human moderation, but rather help

surface points of contention for moderators to further explore and points of agreement to serve as a basis for fairer decision making.

8 Conclusions and Future Work

In this paper, we looked at whether large language models can extract value, value tensions, and consensus points from deliberative transcripts. We explored the effect of using different prompting strategies and models on the results of a metric evaluation and LLM-as-a-judge study.

With respect to SQ1, we tested out 3 different models, Gemma 2, Qwen 2.5 and Mistral, where Gemma 2 achieved the strongest metric performance (F1: 0.540), followed by Mistral and Qwen, although the variance in F1 scores does not indicate a "dominant" performance. LLM-as-a-judge scores were broadly positive, with Mistral achieving the highest average Likert score, although since there was little variance in scores, all 3 models were reasonably capable of the task. With respect to SQ2, we tested out 3 different prompting strategies: zero-shot, few-shot and chain of thought prompting. The two evaluation methods disagree on the best strategy, where the metric evaluation ranked few-shot highest, whereas LLM-as-a-judge ranked CoT the highest, indicating that the two methods captured different qualities of extraction. The metric evaluation ranked CoT the lowest, indicating the lowest accuracy and coverage of extraction due to the complexity possibly being added from the prompt, and the general findings that CoT prompting works better with larger models. Finally, with respect to SQ3, no single model-prompt combination dominated across both evaluation methods, with Gemma zero-shot performing best metric-wise and Gemma CoT performing best with LLM-as-a-judge, which suggest that the interaction between model and prompting strategy is sensitive to how quality is defined and that conclusions about the best combination depend on the evaluation method prioritised, since the rankings between the two methods showed a slight negative correlation.

The goal of this paper was to lay the groundwork and intuition of using LLMs for the extraction of value, value tensions, and consensus points. Though this paper covered the basics of using these techniques for extraction, future work can use higher parameters and stronger models that can increase the reasoning capabilities. The prompt tuning methods can also be refined to branch onto more prompting techniques that enhance reasoning tasks such as these. We can also evaluate the model outputs through a human evaluation study, rather than relying on ground-truth measures or an LLM as a judge, to see how useful these extractions are to actual humans.

Ultimately, this study demonstrates that LLMs show genuine promise for supporting deliberative processes at scale. Surfacing not only consensus points but also underlying disagreements, LLMs can help ensure that the full complexity and nuances of deliberation are effectively handled, rather than flattening conversations into a summary that risks obscuring the values that matter most.

9 Acknowledgments

This research was initially supposed to conduct a within-subjects human evaluation experiment to evaluate the usefulness of the LLM outputs. Ethical approval for human research was applied for before the start of the research project, but due to administrative delays, we did not receive approval in time to conduct the experiment and present the results in this

paper. Hence, the project had to be altered to accommodate the lack of human experimentation in the form of a metric evaluation and LLM-as-a-judge study.

A Annotator Instructions

Overview

You are annotating transcripts of UK House of Commons debates as part of a research project investigating whether large language models (LLMs) can extract deliberative content. Your annotations will serve as ground truth that is used to tune and evaluate LLM prompts. It is therefore important that your annotations are as precise and consistent as possible.

You and the researcher will annotate each transcript independently. Then meet to discuss disagreements if there are any and reach a consensus annotation. The final consensus annotation is what will be used for prompt tuning.

The Three Constructs

1. **Value**

   A value is a stable, abstract principle that guides a speaker's judgment about what is important. It is not a policy position or opinion; it is the underlying principle motivating that position.

   > *Ask yourself: what does this speaker ultimately care about, beyond just winning the argument?*

2. **Value Tension**

   A value tension occurs when two independently legitimate values come into direct conflict over the same issue in the debate. Both values must be present in the transcript; you cannot infer one from silence.

   > *Ask yourself: are two speakers appealing to different values to reach opposite conclusions about the same thing?*

   A value tension requires you to have already identified both values separately. Do not annotate a tension without first annotating both constituent values.

3. **Consensus Point**

   A consensus point is a claim that receives active endorsement from participants who otherwise hold divergent positions. This does not require full unanimity; it requires overlapping agreement across opposing sides.

   > *Ask yourself: is there something both sides appear to agree on, even if they disagree on everything else?*

Annotation Format

Use the provided spreadsheet to fill in the information. The columns are defined as follows:

Column	Definition
Transcript	File name
Speaker	Name of speaker as it appears in the transcript
Verbatim Quote	The shortest excerpt from the transcript that is sufficient to demonstrate the construct
Construct Type	*Value* / *Value Tension* / *Consensus Point*
Label	A short phrase naming the construct (e.g. *"public safety"*, *"safety vs. right to strike"*, *"patient care is paramount"*)
Notes	Optional context notes

B Prompts

B.1 Starting Prompts

Zero Shot Prompt

Identify the values, value tensions, and consensus points present in the following deliberative transcript. The constructs are defined as follows:

Value A stable and abstract principle that guides a person's judgment about what is important.

Value Tension A situation in which two independently legitimate values come into conflict.

Consensus Point A claim or principle that receives endorsement from all, or from a subset of, participants in the discussion.

For each construct identified, provide the following:

Speaker Name of the speaker as it appears in the transcript.

Quote The shortest excerpt from the transcript sufficient to demonstrate the construct.

Construct Type Value / Value Tension / Consensus Point.

Label A short phrase naming the construct (e. g., "public safety", "safety vs. privacy", "human oversight is necessary").

Notes Optional : a brief explanation of why this construct was identified if it is not immediately obvious from the quote.

Few-Shot Prompt

Identify the values, value tensions, and consensus points present in the following deliberative transcript. The constructs are defined as follows:

Value A stable and abstract principle that guides a person's judgment about what is important.

Value Tension A situation in which two independently legitimate values come into conflict.

Consensus Point A claim or principle that receives endorsement from all, or from a subset of, participants in the discussion.

For each construct identified, provide the following:

Speaker Name of the speaker as it appears in the transcript.

Quote The shortest excerpt from the transcript sufficient to demonstrate the construct.

Construct Type Value / Value Tension / Consensus Point.

Label A short phrase naming the construct (e. g., "public safety", "safety vs. privacy", "human oversight is necessary").

Notes Optional : a brief explanation of why this construct was identified if it is not immediately obvious from the quote.

Following are two examples:

Transcript:

> **Speaker A:** We need more surveillance in private and public spaces because the crime rates are through the roof.
> **Speaker B:** But that would infringe on the rights of the residents of the private spaces, surveillance in public spaces is justified though.

Output:

Speaker	**Quote**	**Construct Type**	**Label**
Speaker A	"we need more surveillance in private and public spaces"	Value	Public Safety
Speaker B	"but that would infringe on the rights of the residents"	Value	Privacy
Speaker A / Speaker B	"we need more surveillance in private and public spaces" / "but that would infringe on the rights of the residents"	Value Tension	Public Safety vs. Privacy
Speaker A + Speaker B	"we need more surveillance... public spaces" + "surveillance in public spaces is justified"	Consensus Point	Surveillance in Public Spaces

Transcript:

> **Speaker A:** Due process is too slow, we should weed out suspicious individuals using an automated process.
> **Speaker B:** Due process is integral for a fair democracy, speeding it up with an automated process may introduce bias.

Output:

Speaker	Quote	Construct Type	Label
Speaker A	"due process is too slow"	Value	Efficiency
Speaker B	"speeding it up... may introduce bias"	Value	Bias Prevention
Speaker A / Speaker B	"we should weed out suspicious individuals using an automated process" / "speeding it up with an automated process may introduce bias"	Value Tension	Efficiency vs. Fairness

Chain of thought prompt

Identify the values, value tensions, and consensus points present in the following deliberative transcript. The constructs are defined as follows:

Value A stable and abstract principle that guides a person's judgment about what is important.

Value Tension A situation in which two independently legitimate values come into conflict.

Consensus Point A claim or principle that receives endorsement from all, or from a subset of, participants in the discussion.

Before identifying the constructs, work through the following steps one by one:

Step 1 : Understand the speakers For each speaker, write one sentence summarising what position they are taking and what they seem to care about most.

Step 2 : Identify the values For each speaker, identify the underlying value motivating their position.

Step 3 : Identify the value tensions Look across the values you identified in Step 2. Ask yourself: are any of these values in direct conflict with each other? If so, are both values independently legitimate, or is one clearly wrong? Only flag it as a tension if both are legitimate.

Step 4 : Identify the consensus points Look for moments where speakers with otherwise different positions agree on something, even implicitly. Ask yourself: is there any claim that none of the speakers contest?

Step 5 : Produce the final output For each construct identified in Steps 2, 3 and 4, provide the following:

Speaker Name of the speaker as it appears in the transcript.

Quote The shortest excerpt from the transcript sufficient to demonstrate the construct.

Construct Type Value / Value Tension / Consensus Point.

Label A short phrase naming the construct (e. g., “public safety”, “safety vs. privacy”, “human oversight is necessary”).

Notes Optional : a brief explanation of why this construct was identified if it is not immediately obvious from the quote.

C Prompt Tuning Data

C.1 Metric Instructions

1. For each consensus, the original value would also need to present in the LLM response at least once (deduplicated).
2. Correctly identified labels with incorrect construct types are marked incorrect.
3. A consensus / value tension point must identify all participating speakers and include at least one supporting quote demonstrating the shared position. Individual quotes per speaker are not required provided the identified quote sufficiently demonstrates the convergence.
4. A quote attributed to a speaker who does not appear in the transcript, or a verbatim quote that does not exist in the transcript, will be counted as a hallucination and the construct will be marked as a false positive regardless of label correctness.
5. Labels that are semantically equivalent to the ground truth label will be counted as correct matches.
6. A construct must include a verbatim or near-verbatim quote from the transcript to be counted. Constructs identified without any supporting quote will be marked as false positives.
7. If the model output contains no instances of a particular construct type (e.g. no value tensions identified at all), all ground truth instances of that type will be counted as false negatives.

C.2 Iteration Two Prompts

In order to not repeat the basic prompt templates again, we will only mention the changes made to each prompt.

General changes

These were changes made to all the prompts.

We added : ’Only identify a value tension if you can quote two different speakers invoking conflicting values. Do not infer tensions that are not explicitly present in the transcript.’ after the definition of value tensions.

Added ’A consensus point requires explicit agreement signals between speakers such as "I agree", "I support", or repeated endorsement of the same claim. Do not infer consensus from speakers discussing the same topic.’ after the definition of consensus points.

Added ’Be thorough, identify all instances of each construct type present in the transcript, not just the most obvious ones.’ at the end of the instructions, before the transcript.

Chain of thought prompt

As well as the changes made above, we also decided to reduce the number of steps by merging step 1 and step 2 into : 'Understand the speakers. For each speaker, write one sentence summarising their position and what value is motivating their position.'. This is to reduce the complexity of the steps to improve the overall COT performance.

C.3 Iteration Three Prompts

General Changes

We added : 'Values are often expressed through phrases such as "we/I need", "it is important that", "we/I must ensure", "our/my priority is", "we/I have an obligation to", or any statement where a speaker reveals what principle motivates their position.' after the definition of values.

We added : 'Some transcripts might not contain value conflicts or consensus points in the transcripts.' after the construct definitions.

Few shot prompting

We added the following new example: 'Speaker A: We have an obligation to the public to maintain the highest standards. Speaker B: I agree entirely, though we must also ensure the burden on smaller manufacturers remains proportionate. Speaker C: Both points are well made.

Output: Speaker: Speaker A | Quote: "we have an obligation to the public to maintain the highest standards" | Construct Type: Value | Label: Public Obligation Speaker: Speaker B | Quote: "the burden on smaller manufacturers remains proportionate" | Construct Type: Value | Label: Proportionality Speaker: Speaker A + Speaker B + Speaker C | Quote: "we have an obligation...highest standards" + "I agree entirely" + "Both points are well made" | Construct Type: Consensus Point | Label: High standards are necessary'

Chain of thought prompting

We added the following COT example :

'Transcript: Speaker A: We need more surveillance in private and public spaces because the crime rates are through the roof. Speaker B: But that would infringe on the rights of the residents of the private spaces, surveillance in public spaces is justified though.

Step 1 : Speaker A thinks that crime rates are too high and thus surveillance is needed, they value surveillance. Speaker B thinks that surveillance infringes on the rights of the residents in private spaces but in public spacces it is ok, they value privacy in private spaces.

Step 2 : Speaker A's value of using surveillance in private spaces conflicts with Speaker B's values of surveillance infringing on rights in private spaces.

Step 3: Speaker B agrees with Speaker A that surveillance in public spaces is justified.

Step 4: Speaker: Speaker A | Quote: "we need more surveillance in private and public spaces" | Construct Type: Value | Label: Public Safety Speaker: Speaker B | Quote: "but that would infringe on the rights of the residents" | Construct Type: Value | Label: Privacy Speaker: Speaker A / Speaker B | Quote: "we need more surveillance in private and public spaces" / "but that would infringe on the rights of the residents" | Construct Type: Value Tension | Label: Public Safety vs. Privacy Speaker: Speaker A + Speaker B | Quote:

"we need more surveillance...public spaces" + "surveillance in public spaces is justified" | Construct Type: Consensus Point | Label: Surveillance in Public Spaces'

D Reviewer Instructions

What you are reviewing : You will be given 3 transcripts of simulated citizen assembly discussion on public safety topics. These transcripts are being used in a research study investigating whether AI can extract meaningful content from deliberative conversations. Before they are used, we need to check that they feel realistic and naturalistic.

Your role : You are reviewing whether the transcripts feel like a plausible real conversation.

Instructions :

Go through each transcript and mark any moment where something feels off. Specifically ask yourself:

- Does this feel like something a real person would actually say in a meeting, or does it sound scripted?
- Does any speaker suddenly change their position too easily or too quickly?
- Does any speaker repeat themselves or contradict themselves in a way that feels unnatural?
- Does the conversation flow naturally or are there moments where the topic changes abruptly without reason?

For every violation that you identify please provide :

- Speaker and violating quote
- What specifically feels unnatural

E Evaluation Prompt

You will be given an LLM output of the extraction of value, value tensions, and consensus points from the provided deliberative transcript. The definitions of the constructs are as follows:

Value: A stable and abstract principle that guides a person's judgment about what is important. Value Tension: A situation in which two independently legitimate values come into conflict. Requires quotes from two different speakers invoking conflicting values over the same issue. Consensus Point: A claim or principle that receives endorsement from all, or from a subset of, participants in the discussion. Requires explicit agreement signals between speakers such as repeated endorsement of the same claim or explicit affirmations.

Evaluate the extraction on the following four criteria using a 1-5 Likert scale:

Faithfulness : are quotes verbatim / near verbatim from the transcript and not hallucinated? 1 - most quotes are hallucinated or unverifiable 5 - all quotes are verbatim or near-verbatim from the transcript

Construct Correctness : do extractions match the definitions and are correctly labeled? 1 - most extractions violate the construct definitions 5 - all extractions correctly match their definitions

Coverage : does the output identify all major constructs present, or does it miss obvious ones? 1 - many obvious constructs are missed 5 - all major constructs are identified

Label Quality : Are labels appropriately abstract summaries of the construct rather than paraphrases of the quote? 1 - labels merely restate the quote 5 - labels are concise and accurately named

Transcript: transcript

Response: response

Respond in the following JSON format only, with no preamble: "faithfulness" : "score" : X, "justification" : "...", "construct_correctness" : "score" : X, "justification" : "...", "coverage" : "score" : X, "justification" : "...", "label_quality" : "score" : X, "justification" : "..."

F Model Parameters

Seed	Temperature	top_p	top_k
42	0.0	1.0	1

Table 17: Model Parameters for Extracting Constructs from Training and Test Transcripts

Seeds	[7, 13, 42, 67, 69, 99, 111, 162, 404, 909]
Temperature	0.2

Table 18: Model Parameters for LLM-as-a-judge

References

[1] Dhruv Agarwal, Farhana Shahid, and Aditya Vashistha. Conversational agents to facilitate deliberation on harmful content in whatsapp groups. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–32, November 2024.

[2] Ibukun Babatunde, Obiabuchi Nnanna, and Mark Klein. Moderating large scale online deliberative processes with large language models (llms): Enhancing collective decision-making., March 2025.

[3] Nilesh Barla. What is self-consistency prompting?, March 2026.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[5] Martin Carcasson and Leah Sprain. Beyond problem solving: Reconceptualizing the work of public deliberation as deliberative inquiry. *Communication Theory*, 26, April 2015.

[6] Simone Chambers. Deliberative democratic theory. *Annual Review of Political Science*, 6(Volume 6, 2003):307–326, 2003.

[7] Cheng-Han Chiang and Hung yi Lee. Can large language models be an alternative to human evaluations?, 2023.

[8] Katherine M. Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E. Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua B. Tenenbaum, and Thomas L. Griffiths. Building machines that learn and think with people, 2024.

[9] Xia Cui, Ziyi Huang, and Naeemeh Adel. Bias in, bias out: Annotation bias in multilingual large language models, 2025.

[10] Suyash Fulay, Dimitra Dimitrakopoulou, and Deb Roy. The empty chair: Using llms to raise missing perspectives in policy deliberations, 2025.

[11] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.

[12] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298, 2005.

[13] Julia Jaremko, Dagmar Gromann, and Michael Wiegand. Revisiting implicitly abusive language detection: Evaluating llms in zero-shot and few-shot settings.

[14] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[15] Sam Kaner, Jamie Watts, and Emile Frison. Participatory decision-making: The core of multi-stakeholder collaboration. *Institutional Learning and Change (ILAC) Initiative, ILAC Briefs*, 01 2008.

[16] Georgi Karadzhov, Andreas Vlachos, and Tom Stafford. The effect of diversity on group decision-making, 2024.

[17] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023.

[18] Molly Murton, Ellie Boulton, Shona Cross, Ambar Khan, Swati Kumar, Giuseppina Magri, Charlotte Marris, David Slater, Emma Worthington, and Elizabeth Lunn. Harnessing large-language models for efficient data extraction in systematic reviews: The role of prompt engineering. *Cochrane Evidence Synthesis and Methods*, 3(6):e70058, 2025.

[19] Soya Park, Hari Subramonyam, and Chinmay Kulkarni. Thinking assistants: Llm-based conversational assistants that help users think by asking rather than answering, December 2023.

[20] John Rawls. The idea of an overlapping consensusâ . *Oxford Journal of Legal Studies*, 7(1):1–25, March 1987.

[21] Shalom H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 25 of *Advances in Experimental Social Psychology*, pages 1–65. Academic Press, 1992.

[22] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, 2024.

[23] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, October 2024.

[24] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges, 2025.

[25] Chao Wei, Yanping Chen, Kai Wang, Yongbin Qin, Ruizhang Huang, and Qinghua Zheng. Apre: Annotation-aware prompt-tuning for relation extraction. *Neural Processing Letters*, 56, February 2024.

[26] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824â24837, December 2022.

[27] Vasiliki Xiromeriti. What is collective deliberation? collective deliberation as shared reasoning. *Episteme*, pages 1–15, March 2025.

[28] Mohammad Amin Zadenoori, Vincenzo De Martino, Jacek Dabrowski, Xavier Franch, and Alessio Ferrari. Does model size matter? a comparison of small and large language models for requirements classification, 2025.

[29] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. Deliberating with ai: Improving decision-making for the future through participatory ai design and stakeholder deliberation. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–32, April 2023.

[30] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models, 2024.

[31] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.