How to design post-reflection dialogue from transcripts using the identified values, value tensions, and consensus points?

A deliberative approach to modeling retrospection ex post facto in multi-stakeholder decision-making scenarios.

Victor-Emanuel Clatici[1]
Supervisor(s): Willem-Paul Brinkman[1], **Michael 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: Victor-Emanuel Clatici
Final project course: CSE3000 Research Project
Thesis committee: Willem-Paul Brinkman, Michael Grauwde, Sole Pera

Abstract

Large Language Models (LLMs) excel at natural language tasks, yet most contemporary systems and tool prioritize providing an answer over fostering reflection and deliberation. This research investigated whether LLM-based tools can generate post-reflection dialogue in multi-stakeholder decision making scenarios by using identified value tensions and points of contention found in transcripts.

A Deliberative AI approach was developed using publicly available transcripts and several open-sources LLMs. The generated reflective dialogue was subsequently evaluated through Synthetic Personae evaluators according to the five metrics established: safety, privacy, autonomy, societal well-being, and points of contention. Two different prompting strategies: single-turn and multi-turn were deployed to see if there were meaningful differences between the two.

The results indicated that the methodology can produce reflective dialogues that are perceived positively, exceeding the predefined success threshold. Furthermore, the iterative multi-turn interactions were found to improve perceived satisfaction compared to the single-turn approach on average.

Although limited to English language deliberations, the findings demonstrate the feasibility of using Deliberative AI to support reflection and, rather than proposing a universal solutions, this work provides a reproducible proof of concept that can be adapted based on future models, transcript contexts, and languages, motivating the development of more LLM-based deliberative systems.

1 Introduction

Large Language Models(LLMs) and their capabilities have quickly evolved in recent times, from the humble beginnings of perceptrons to the massive architecture that seems to creep little by little into the everyday lives of everybody[20]. But with quick advancements in the computational power that systems present in the modern age another question comes: is their coupling with the human element done responsibly? Are Large Language Models and Conversational Agents a silver bullet that can be applied indiscriminately to each and every single problem or difficulty that humanity is facing? While this may be an alluring idea, the literature shows this cannot be boiled down to such an easy answer.[5, 21, 35, 38] Other work has been done on the broader topic of development of LLM-based deliberative agents[16, 18, 24, 32, 39], as opposed to the more widely encountered explainable AI (XAI), as solutions that could be deployed to achieve better results in more social scenarios, or in situations that do not hold clear, objective answers and which would rather encourage the strengths of a deliberative approach as opposed to the shortcomings that XAI and other techniques present.

Work on the subject of Artificial Intelligence, broadly, has accelerated in the years following OpenAI's success with their ChatGPT chatbot[20]. However, there is no currently available solution that works for every use case [21]. A lot of actionable work on the subject has been carried out, as such, in the recent period of three years. Deliberative Dynamics and Value Alignment have been researched[32] and the impact that a thinking assistant as opposed to a proverbial oracle would provide has also been investigated[24] in order to see the mental impact that such tools bring about on the user. The effects that a deliberative agent could have on reflections about physical[16] or mental activities [39] has also begun to explored. One of the more low level initiatives towards the implementation of deliberative agents [18] also advocates for a paradigm shift in tools used for certain tasks, citing the pitfalls of current LLM-based AI tools and chatbots: namely that overreliance on current

solutions can lead to users becoming compromised in their own mental faculties[6, 10, 27], by overrelying on the results or processing them with less scrutiny than would be required. The results of the studies indicated the existence of areas of improvement within the LLM-based solutions deployed today as well the shortcomings of the proposed solutions. It has been documented that current AI-assisted decision-making systems give an answer as a whole instead of a process that can be questioned or improved: the answer that is given has to either be accepted or rejected[18, 28]. This is, naturally, a issue that encourages either conformity with the system or direct distrust if the user notices habitual incorrect answers, raising the need for the development of a alternative to the current solutions is further encouraged.

Concretely, this research explored the following research question in order to contribute towards the development of an LLM-based solution that can address the gap presented by conventional solutions in these scenarios.

- RQ: How to design post-reflection dialogue from transcripts using the identified values, value tensions, and consensus points?

In order to understand whether it is possible to construct reflective dialogue by means of a Large Language Model or LLM based tools, the problem was split into two sub Research Questions:

- SRQ1: Is it possible to design post-reflection dialogue from transcripts using the identified values, value tensions, and consensus points using a Large Language Model?
- SRQ2: How does a single-turn prompting strategy compare with a multi-turn strategy in in designing post-reflection dialogue?

The first sub research question explored whether reflection can be accomplished with the method to be described: the feasibility of the research question; the second sub research question explored whether there are noticeable differences between a multi-turn approach and a single-turn approach, the technique use for the first sub research question.

The rest of the paper contains the methods by which the Research Question was answered, background and related work, and precise problem descriptions. They are followed by the presentation of the results, a discussion and conclusion about them, as well as the reproducibility of the experiments. Additionally, an Appendix is present with 2 more additional Tables.

2 Background and Related Work

In order to properly begin describing the methods we employed and the results, it is important to establish the baseline definitions for contentious concepts as well as the importance of this research. When we talk about deliberation within the context of this research, we are referring to "communication that induces reflection on preferences, values and interests in a non-coercive fashion"[19]. A Large Language Model, hereafter shortened to LLM, is a "large-scale, pre-trained, statistical language model based on neural networks" [20], whereas Conversational Agents are "programs that are designed to communicate with users using natural language" [15].

2.1 Explainable AI or Deliberative AI

Conventional LLMs and Conversational Agents are usually tuned for the delivery of a definitive answer to a question that the user has, or in applying tools to solve a specific task, when given specific access[20, 22]. Explainable AI techniques are further employed to varying degrees within these models and beyond in order to elucidate from the "Black Box" aspect that several models present[1]. Other techniques such as mimicking human conversational patterns have also proven to be effective in giving these models wider use cases[24], as well as serve as important pillars upon which users can more easily converse with the respective models. While the anthropomorphism deployed at large has the objective of making the models more "familiar" to the user, they can have the opposite effect if deployed poorly or in inappropriate situations[27]. In short, while conventional models employ varying techniques in order to elucidate their thinking patterns, they are not perfect. To this end, the concept of Deliberative AI was introduced and is defined as such: "assistants powered by large language models designed to enable Human-AI Deliberation-a collaborative process where humans and AI deliberate together to resolve conflicting perspectives in decision-making tasks—by seamlessly integrating domain-specific models with large language models (LLMs)" [18].

Deliberative AI, as its name implies, was conceived in order to bridge the computational power and possibilities that LLMs present with the human component that can become disillusioned with the ever-present "take it or leave it" answers that LLM-based models tend to favor that erode human agency from the process[6, 18, 28]. This is due to the fact that conventional solutions, even ones that employ XAI techniques as mentioned earlier still suffer from the inherent limitations that most LLMs possess: they are designed around giving an answer, along with a reasoning process occasionally.[10, 28] Deliberative AI serves as a foil to conventional solutions, with the aim of implicating the user within the process of finding a answer, but not necessarily the answer as other solutions do. As such, Deliberative AI aims not to regurgitate an answer that the user may not understand, but rather to engage and challenge them in order to arrive at solution together, promoting their critical thinking[24, 39].

In short, traditional AI architecture, extended by XAI, attempt to make the user understand how the system arrived at a specific answer, but the possibility in engaging with such systems in reflective dialogue is limited by their underlying architecture: they are suited to find an answer and subsequently explain the method by which they arrived at it. Deliberative AI, on the other hand, practices the inverse: it engages with the user, creating the method through which an answer, but not the answer is reached. This design puts the user in the proverbial driver's seat instead of the passenger's when engaging with the model and as such it was selected as the paradigm to be explored further in the research.

2.2 Synthetic Personae as evaluators

During the research, the approval for human elicitation was not obtained, and as such this research needed to find a way to substitute human input for the purposes of evaluating the dialogue that was created. The usage of Synthetic Personae or similar variants of LLM-as a-judge is a technique not uncommon in LLM-based research[2, 23, 34]. The solution that we arrived as was to deploy a solution to simulate human decision making and perceived emotions as closely as possible, while still being ethical. To that end, Synthetic Personae, as explored by Batzner et alia. [4] were chosen as the method to bridge the gap between solution and confirmation. The work of Sanchez [34] also provided motivation for the multi-turn

approach described in the Method section, with the recommendation of "proposing more challenging mental capacities or phenomena to simulate", which in this study manifested as the more demanding process of having a longer conversation, compared to just the single-turn approach. The "Persona Transparency Checklist" introduced by Batzner et al. which was built on top of previous attempts was invaluable in confirming that the inconsistencies observed within their research would not be repeated within this one[4]. Further details regarding the deployment of the Synthetic Personae are available within the method section of the paper.

The crux of this research can then be summed as such: we explored the problem of creating self-reflection dialogues with LLM tools in order to find a technique that proved effective in said scenarios, but which is not necessarily the most effective. In other words, this paper's purpose was to provide a proof of concept of how such a technique can be devised, and possibly be extended, along with the design choices that were taken to bring about its deployment.

In order to know what criteria to use for the evaluation of the created dialogue's perceived satisfaction, four values were identified within the conversations, along with a fifth dimension of points of contention for the purposes of converting the evaluator's assessment into a numeric form. In particular, Safety was defined as "the condition of being protected from harm (or other nondesirable outcomes) caused by non-intentional failure of technical, human or organisational factors"[36], Privacy as "safeguarding the spontaneous, independent, and uniquely individual aspects of the self"[29], Autonomy "refers to the ability of persons to create their own identity and in this way to define themselves."[29], and Societal well-being refers to how an individual feels accepted or welcome in a society or community[33].

3 Method

The method by which the sub research questions were answered comprised the generation of reflective dialogue based on a given transcript and subsequent evaluation of the perceived satisfaction across the four values and points of contention. In the single-turn approach the reflective dialogue was generated and directly evaluated by the Synthetic Persona evaluator. In the multi-turn approach a longer conversation consisting of "turns" took place for the purposes of evaluating whether a longer conversation where both sides engage with the other during the reflective process would yield different results compared to the single-turn approach, with the results being evaluated by the Synthetic Persona after the final message. The evaluations from both approaches were then analysed and interpreted in order to answer both of the sub research questions.

In this section, the methods applied for the realization of this research are further detailed. In order to reach a state in which we could determine whether reflection strategies were possible to begin with, the first subsection comprises the definition of Synthetic Personae that evaluated the prompted dialogue generated within the second subsection, establishing the skeleton for the experiment setup. The third subsection justifies the choice of LLMs deployed, while the fourth subsection explains how they were deployed. The final part subsequently details how the multi-turn approach different from the single-turn.

3.1 Defining the Synthetic Personae to be used

In order to evaluate the reflective dialogue created, Synthetic Personae, detailed by Batzner et al. were deployed[4]. The work references five archetypes commonly used by researchers when carrying out experiments using them. The first and second forms are the "I am" and "You are" Personae respectively. They are often associated with role-playing capabilities. The third form is the "Preferences" form which is centered around prompting of the user on what the Personae is supposed to represent. The fourth form is the "Real Conversations" form based on chat data. They are derived from conversation data and context rather than explicitly describing sociodemographic attributes. The final form is the "Survey Responses" form that builds the Persona with the help of survey responses.

The approach that was taken for the evaluator Personae within the rest of the research was the "Real Conversations" format. The reasons are numerous. Firstly it is an inexpensive way of constructing Personae that are relevant to each scenario[4]. Secondly due to being based on the conversation data that the reflection dialogues are built upon to begin with the intention was for the Personae to be more attuned to the state of mind that the interlocutors had in that particular context. In a manual elicitation scenario people could answer untruthfully[17, 25]. Lastly it was also selected due to the belief that it is the most conducive to reproducibility given that it does not rely on manual creation of the Personae, as opposed to the "I am", "You are", and "Preferences" forms or the "Survey Responses" form which is invalid due to the research not securing human participants.

In order to ensure the correctness of the application of Synthetic Personae, the "Checklist for Persona-based LLM Research"[4] was leveraged. More concretely, the checklist comprises six evaluation metrics that the researchers used in order to ascertain whether a study has developed and used Synthetic Personae in a "transparent, reproducible, and rigorous way". The "Application" dimension has been discussed within this paper within the background and introduction sections. Additionally, the task definition and context were also used within the prompting of the Synthetic Personae, detailing their situation, and task definition. The "Population" dimension addresses the target demographic for the evaluation but given the episodic nature of the lifespans of the personae and their tailoring specifically to the context from which they were created. As such, the population dimension refers to each interlocutor that was supplanted by a Synthetic Personae evaluator, with the common denominator being the provenance of their discussions from within the same transcript context. The "Data Source" dimension is described within the next sub section, where the transcripts that were used are discussed. The "Ecological Validity" dimension is ensured due to the usage of the entire transcript in order to lend more credence to the Synthetic Persona implementing the "Real Conversations" form. The "Reproducibility" dimension of the research is further detailed within this section and within the "Responsible Research" section.

Lastly, a pipeline was established for the "core" of the evaluation aspect of the research: the LLM/LLM-based tool received a transcript from which it subsequently created a reflective dialogue. The reflective dialogue was then given to the Synthetic Persona, inhabited by an LLM as well, and evaluated based on the values mentioned earlier in the paper: safety, privacy, autonomy, societal well-being. Prompting techniques were deployed in order to anonymize the identities of the interlocutors within the transcript, as well as to reduce variance or bias that can arise from their names, given that such bias has been shown within previous research.[26]

3.2 Database Selection and Prompting

The next step involved selecting transcripts for the experiment. A dataset from a bounded timeframe was required, sufficiently narrow to maintain contextual coherence while broad enough to avoid overfitting to a single event or discourse context [7]. Given the focus on deliberative AI in multi-stakeholder decision-making for public safety, the dataset was selected from a setting representative of such deliberative scenarios.

Given the focus on deliberative AI in multi-stakeholder decision-making for public safety, the dataset was selected to reflect a structured and institutionally grounded form of deliberative discourse. To this end, transcripts from the United Kingdom House of Commons were used, as this setting provides a well-established example of formal parliamentary deliberation, characterised by procedural rules, multi-party participation, and documented argumentative exchange.

This constrained selection enables analysis of deliberative dynamics within a consistent and clearly defined context. The research is situated within this setting to examine whether deliberation-enhancing effects can be observed under controlled conditions. At the same time, it is acknowledged that such effects are likely to be sensitive to cultural norms, institutional structures, linguistic variation, and model-specific behaviour, which may influence generalisability across other contexts.

Accordingly, the scope of the research is intentionally bounded to support systematic evaluation within a single deliberative environment. If the approach yields meaningful results in this setting, it provides a basis for further investigation across additional contexts and domains.

The chosen month for the experiments in the first sub question is the month of February 2026, totaling eight days of deliberation from which transcripts during deliberation periods were copied and anonymized. The next step after the selection of the database was to analyze the transcripts within the broader context of their discussion, to later apply the gleaned insights to the LLM or LLM-based tool for efficient prompting of a reflective dialogue. Upon realization of an understanding of the transcripts at large, a strategy was drafted based on prompting techniques that would separate subjectivity from the prompt as much as possible and, as mentioned earlier, the experiment was made to uphold the "Persona Transparency Checklist"[4]. The evaluator Synthetic Persona was likewise given a prompt that incorporated the aforementioned generator reflective dialogue, as well as instructions to evaluate its effectiveness based on the Persona it was adopting, as well as the four defined values: safety, privacy, autonomy, and societal well-being. A successful reflective dialogue strategy was defined as one whose average perceived satisfaction exceeded 50 percent across ten iterations. This threshold was chosen because the objective was not to establish strong consensus, but merely to determine whether the strategy exhibited a net positive effect, thereby serving as a proof of concept.

The final forms of the prompt for the single-turn setup for the reflective dialogue and Synthetic Persona creation and evaluation respectively were:

- Reflective(Generator) Prompt: "You are receiving a transcript of a dialogue in a deliberation within the United Kingdom's House of Commons. Based on the dialogue presented, you are to output a reflection relating to Person X that refers to each value tension and points of contentions that occurred in the transcript related to Person X. These are the following four values. [VALUE DEFINITION]. The following is the transcript [TRANSCRIPT]. You are to address Person X and generate the reflection

accordingly in an exploratory manner towards Person X about whether the four values were respected.".

- Synthetic Persona creation and evaluation: "You are receiving a transcript of a dialogue in a deliberation within the United Kingdom's House of Commons. Based on the dialogue presented by Person X, you are to process the following dialogue from the perspective of Person X and judge it based on the following four values and how well they were captured by the dialogue in relation to Person X. There are the following four values [VALUE DEFINITION]. The following is the transcript [TRANSCRIPT]. The following is the reflection dialogue [DIALOGUE]. Output five metrics, each from 1-10 based on how you perceived the effectiveness with each of the four values and one for the points of contention respectively were explored in relation to the transcript.".

The Transcript, Value Definitions, and Dialogue were represented in natural language using raw text files without additional formatting or augmentation. The results across the selected days of deliberation were aggregated for subsequent analysis.

The prompt design was informed by prior work on structured and instruction-based prompting. Wei et al. [37] demonstrate that structured prompting and explicit decomposition of reasoning steps improve performance on complex tasks. Reynolds and McDonell [30] show that instruction-style prompts can influence model behaviour beyond few-shot examples, indicating that prompt structure is a relevant factor in shaping model outputs.

3.3 Choice of Large Language Models

In order to evaluate whether the proposed methodology for generating post-reflection dialogue can be applied consistently across different contemporary large language model families, three models were selected for the experiments: Llama 3.2 3B, Gemma 4 E2B, and Qwen 3 V1 4B.

The choice to employ multiple models instead of relying on a single solution was motivated by the desire to reduce the likelihood that the observed results are artifacts of a particular architecture, training corpus, or design philosophy. The selected models originate from different development ecosystems and exhibit differing strengths with respect to instruction following and reasoning. Their use therefore provides an opportunity to evaluate the robustness of the proposed methodology across heterogeneous LLM families rather than optimize for the characteristics of a specific model [20, 38].

The selected models represent distinct approaches to language modelling and reasoning while maintaining sufficient capability to perform the tasks required by this research. Their inclusion allows the experiments to evaluate whether the proposed prompting strategies exhibit consistent behaviour across heterogeneous model families. Consequently, the purpose of model selection is not to maximize performance, but to determine whether the reflective dialogue framework demonstrates robustness when applied to multiple contemporary Large Language Models [20].

Consequently, the model selection process prioritizes external validity and methodological robustness over maximizing the performance of any individual system. The experiments therefore seek to determine whether the proposed prompting strategies exhibit consistent behaviour across different model families rather than identify a universally optimal model.

3.4 Deployment of Large Language Models

To minimize model specific bias in both the generation of reflective dialogue and its subsequent evaluation, the roles of the Large Language Models were deliberately separated and decoupled across tasks. In this pipeline, one model instance is used exclusively for generating the reflective dialogue from the input transcript, while a separate model instance is used to instantiate the Synthetic Persona responsible for evaluation. This separation was made with the intention of reducing the the risk that artifacts of a single model's training distribution, alignment behaviour, or stylistic tendencies influence both the construction and assessment of the same output.

All models were prompted using a fixed evaluation schema, identical value definitions, and consistent temperature settings to further constrain variance arising from prompting differences. By enforcing role separation, cross-model rotation, and standardized prompting conditions, the pipeline aims to isolate the effect of the reflective dialogue strategy itself, rather than confounding it with idiosyncratic behaviours of a particular model family.

In order to balance the reproducibility of the research with the inherently stochastic nature of LLM generation, all experiments were conducted using a fixed temperature value of 0.3 and identical decoding parameters, while varying the random seed across iterations. A low temperature was selected to constrain excessive output diversity and ensure that differences in observed results arose primarily from the prompting strategies rather than from sampling noise. At the same time, varying the seed between runs allowed multiple independent generations to be obtained and aggregated, capturing the natural variability present in the models without compromising experimental consistency. This approach provides a compromise between deterministic evaluation and the need to account for the probabilistic behaviour of modern language models.

3.5 Comparison between Single-Turn and Multi-Turn Strategies in Creating Reflective Dialogue

Having established the feasibility of generating and evaluating reflective dialogue, the second stage of the research focused on comparing different interaction strategies. More specifically, the objective was to determine whether a single-turn approach or a multi-turn approach would be more effective in facilitating reflection and increasing perceived satisfaction with respect to the values defined earlier. The evaluation framework remained largely identical to that employed for the first sub research question. The transcript selected for analysis was provided to an LLM, which generated an initial reflective dialogue based on the identified values, tensions, and consensus points. This output was then presented to the Synthetic Persona evaluator, which assessed the reflection according to the criteria previously defined.

For the single-turn strategy, the generated reflection was treated as the final output and evaluated directly by the Synthetic Persona. In this case, no additional interaction occurred and the evaluation concluded after a single exchange.

For the multi-turn strategy, the generated reflection was instead treated as the initiation of a dialogue. After receiving the reflective prompt, the Synthetic Persona produced a response corresponding to the perspective it embodied. This response was then provided back to the LLM, which generated a continuation of the conversation. The process was repeated iteratively, allowing both entities to engage in a sequence of exchanges intended to further explore the identified values and tensions, a strategy previously found to be effective[24].

Deliberation does not necessarily seek unanimous agreement or a single objectively correct answer, but rather a state in which participants have sufficiently explored the issues at hand and reached a level of satisfaction with the outcome of the discussion. As argued by Mansbridge et al. [19], deliberative processes must account for the role of self-interest and conflicting perspectives, implying that successful deliberation is not contingent upon complete consensus. Consequently, the interaction terminated either when the Synthetic Persona explicitly indicated satisfaction with the discussion or when a predetermined maximum number of turns had been reached. In this research, the number of turns used was four for both sides of the conversation. Due to the prompts differing slightly to allow for the possibility of the longer dialogue, the full differences comprised the adaptation of the prompts into the starting, interlude, and final prompts throughout the conversation.

The design of the multi-turn strategy was inspired by the mini-dialogue approach employed by Kocielnik et al. [16], which utilized follow-up questions to encourage reflection. In the context of this research, the concept was extended from isolated follow-up questions to a sustained exchange between the reflective dialogue generator and the Synthetic Persona. The purpose of this modification was to investigate whether introducing iterative interaction and allowing the evaluator to actively participate in the conversation would result in a more effective reflective process than presenting a single, static response.

Both strategies were evaluated under identical conditions using the same transcripts, prompting techniques, temperature settings, and Synthetic Persona configurations in order to isolate the effect that the interaction strategy itself had on the perceived satisfaction of the values under consideration.

4 Results

This section details the experiment results that were obtained based on the aforementioned method. In total, the results are aggregated based on the 480 Single-Turn and the 480 Multi-Turn conversations that were conducted. Each pair of LLMs operated on the same seed range of 42-51 on each transcript sequentially in order to ensure variance and reproducibility in the results, with a temperature of 0.3, Top K sampling of 40, Repeat Penalty of 1.1, Top P Sampling of 0.95, and Min P Sampling of 0.05. The following statistics presented for each Research Question were made by aggregating the obtained evaluation results from the Synthetic Personae across the entire aggregate, with the Appendix holding additional tables that list the obtained evaluations for each value individually.

4.1 SRQ1: Possibility of Reflective Dialogue Creation

In Table 1 are listed the aggregate evaluations produced in the Single-Turn strategy by the Synthetic Personae, rated on a scale from 1 to 10

Table 1: Single-Turn Aggregated Results across all Transcripts across all five metrics

Generator / Evaluator	Llama 3.2	Gemma 4	Qwen 3 V1 4B
Llama 3.2	N/A	7.6725	8.1750
Gemma 4	7.0475	N/A	7.6500
Qwen 3 V1 4B	5.7450	7.3475	N/A

The results obtained for the single-turn strategy indicate that the approach was suc-

cessful according to the criteria established earlier in the research. Across all model combinations, evaluation metrics, aggregated over all eight transcripts, the average perceived satisfaction consistently exceeded the predefined threshold of 50 percent over ten iterations. Since this threshold was chosen to represent a net positive effect rather than strong consensus, the results suggest that even a single, static reflective dialogue is capable of capturing value tensions and points of contention in a manner perceived favorably by the Synthetic Persona evaluators. Consequently, the findings support the feasibility of employing single-turn prompting as a proof of concept for the construction of reflective dialogue in multi-stakeholder decision-making scenarios. While the results do not imply that the generated reflections are optimal or universally applicable, they did demonstrate that the approach is sufficiently effective to warrant further investigation and comparison with more interactive strategies. We can also observe that among the obtained results, privacy was rated the lowest on the aggregate, while societal well-being was narrowly higher than the effectiveness with which points of contention were explored. For the extent of the research question however, these results showed the feasibility of the methodology in making effective reflective dialogue, and as such answered the first sub research question. It is important to note, however, that not all model combinations exceeded the 50 percent threshold for each of the criteria, namely the combination of Qwen 3 V1 4B as generator and Llama 3.2 as evaluator. In that specific scenario, the average rating for the Privacy value was 4.1375 across the eighty iterations(see Figure 3 in Appendix A).

4.2 SRQ2: Comparison between Single-Turn and Multi-Turn Approaches

In Table 2 are listed the aggregate evaluations produced in the Multi-Turn strategy by the Synthetic Personae, rated on a scale from 1 to 10.

Table 2: Multi-Turn Aggregated Results across all Transcripts across all five metrics

Generator / Evaluator	Llama 3.2	Gemma 4	Qwen 3 V1 4B
Llama 3.2	N/A	8.04305	6.8350
Gemma 4	8.0700	N/A	8.6350
Qwen 3 V1 4B	7.8325	8.7130	N/A

The satisfaction metrics obtained from the multi-turn strategy were perceived higher in the total average scores for each value compared to the single-turn approach. It can, however, also be seen that a combination of LLMS, namely Llama 3.2 as generator and Qwen 3 V1 4B as evaluator, had lower scores in the multi-turn approach, compared to the single-turn one.(see Figures 3, 4 in Appendix A). It can also be observed that whereas in the single-turn approach there was one metric for the combination of Qwen 3 vV1 4B/Llama 3.2 that did not pass the 50 percent threshold, in the multi-turn approach all of the ratings for a specific value or points of contention were, in the aggregate, above the threshold. Furthermore, strictly based on numerical values, all of the averages for each metric were higher in the multi-turn approach compared to the single-turn approach. For the extent of the second research question however, these results showed that there are significant differences between the ratings of the single-turn and multi-turn approaches however it also showed a general lower score across both approaches for the privacy value rating, with it being the lowest on the aggregate for both methods. Additionally, highest rated value being societal well-being within both approaches can also be observed.

Taken together, the first sub research question was answered with the described method being successful in generating reflective dialogue that was perceived as being helpful in regards to exploring the values of an interlocutor, while the second sub research question explored the influence that the length of the deliberative dialogue could have on the perceived satisfaction of the evaluator. As such, both of them served towards the answering of the main research question through offering successful examples of post-reflection dialogue and a comparison of different techniques.

5 Discussion and Future Work

5.1 SRQ1: Possibility of Reflective Dialogue Creation

As stated in the Results section, the average perceived satisfaction consistently exceeded the predefined threshold of 50 percent on the aggregate of all transcripts. Since this threshold was chosen to represent a net positive effect rather than strong consensus, the results suggest that even a single, static reflective dialogue is capable of capturing value tensions and points of contention in a manner perceived favorably by the Synthetic Persona evaluators. Consequently, the findings supported the feasibility of employing single-turn prompting as a proof of concept for the construction of reflective dialogue in multi-stakeholder decision-making scenarios.

With the exception of the Qwen 3 V1 4B as generator and Llama 3.2 as evaluator, all of the aggregate ratings were above the 50 percent threshold. This could indicate either an inherent weaker ability for deliberation in this specific combination of models, that the dataset was better suited to the features of the other models, or that the difference arose more from the ordering of the models, rather than the combination itself. When analyzing the results between Qwen 3 V1 4B and Llama 3.2 as a generator/evaluator and evaluator/generator pair respectively we can observe that the evaluator/generator pair outperformed its counterpart on all metrics except for points of contention(see Figure 3 in Appendix A). The differences between ordering of generator and evaluator are present in all of the combinations, and as such indicate that the evaluation was not only impacted by the choice of generator alone or choice of evaluator alone, but that both the pairing and the ordering mattered. This inference is further supported by the Figure 4 of the multi-turn approach where there are noticeable differences between model pairs based on position. The individual per model pair ratings for each value also vary, with the biggest differences being found in the Qwen 3 V1 4B/Llama 3.2 pairing mentioned earlier.

The lowest rated value was privacy while the highest rated was societal well-being, a fact further reflected within the multi-turn approach results as well (see Figure 3 in Appendix A). Without exploring a different dataset it is difficult to determine whether this is an inherent feature of the method that was chosen for the evaluation or of the transcripts themselves.

Overall, the first sub research question addressed the feasibility of the approach by determining whether Large Language Models are capable of generating such reflective dialogue from transcript data and with the results obtained.

5.2 SRQ2: Comparison between Single-Turn and Multi-Turn Approaches

The second sub research question concerned the implementation of reflective dialogue, comparing single-turn and multi-turn prompting strategies in order to determine how reflective

dialogue can be structured most effectively. To this end, several differences can be noticed when comparing the results. As shown in Table 1 and Table 2, we can observe that the Synthetic Personae, on average, rated the perceived satisfaction of the conversation higher across all of the five dimensions when aggregating over the entire transcript dataset.

The Qwen 3 V1 4B/Llama 3 pairing no longer is below the 50 percent threshold in its privacy evaluation and overall all of the model pairings received higher ratings in all average value aggregates compared to the single-turn approach(see Figure 4 in Appendix A). This might indicate either that it is the length of the conversation which influenced the satisfaction metrics or due to the interactive nature of taking turns enabling both sides of the conversation to update their side of the discussion accordingly. This is a feature that was not present in the single-turn approach at all, given that the amount of turns there was the equivalent of one.

The Llama 3.2/Qwen 3 V1 4B alone had lower scores in the aggregate in the multi-turn approach compared to the single-turn when compared to its peers. This result can be due to a model inherent mismatch between the two where this specific combination yields less useful reflective dialogue or it can be caused by context of the transcripts used. In other words, it is worth exploring other datasets or other context to see if similar behaviour might be found in other LLM combinations or if this was an isolated case.

A similar explanation might be applicable to another finding: both within the single-turn and multi-turn approaches the privacy and societal well-being were the lowest and highest rated, respectively, on the aggregate. As stated earlier, a possible cause of this behaviour could be the transcripts themselves. It is possible that certain values were overrepresented or underrepresented within the discussions and as such forming deliberative dialogue over one of the two would prove to be easier to accomplish. Another possible explanation would be in the prompts themselves. It could be that despite the intention of having the language be as neutral as possible, the ordering of the values within the prompt itself might have influenced the dialogue and subsequent evaluation. Another possibility is that LLMs themselves, or at least the selection of the three chosen within the study, find it harder to capture the privacy value and deliberate over it compared to the societal well-being.

The trend that was present throughout the multi-turn approach was, as stated, that of an increase in all evaluation metrics. This increase, based on the aforementioned reasons and discussions, could be caused as a result of inherent LLM attributes, implicit prompting bias, transcript specific differences, or innate differences in how a values is interpreted or understood by itself. While the results did show an increase in the perceived satisfaction we cannot say with certainty that the approach of multi-turn dialogue is intrinsically superior to the single-turn approach. It warrants more research.

As a whole, we answered the research question by providing the proof of concept for the creation of post-reflection dialogue and offering a comparison of two approaches. However, certain things should still be considered: the results do not imply that the generated reflections are optimal or universally applicable, but they did demonstrate that the approach is sufficiently effective to warrant further investigation and comparison with more interactive strategies, such as the multi-turn approach.

5.3 The Usage of Synthetic Personae

Due to the limitations on the TUDelft's part regarding ethics approval for human elicitation, in order to evaluate the effectiveness of the dialogues made based on the different strategies we used Synthetic Personae in order to simulate human testers. While less human in the

literal sense they came with the advantage of easily being populated or tuned to a specific demographic that needs to be involved within a specific experiment or research[4] but with the tradeoff of needing close scrutiny and interpretation to make sure that they are used responsibly[34].

5.4 Bias within the dataset

It is not possible to fully account for all sources of bias and subjective interpretation that may arise from the methodological choices employed in this research. The transcripts selected for analysis were drawn from a randomly selected month of United Kingdom House of Commons debates, with the intention of providing a coherent and bounded dataset for the study of deliberative interactions. However, this selection does not guarantee that the observed behaviours would generalise to other political or deliberative systems, nor to multilingual contexts, where differences in institutional norms, linguistic structure, and cultural framing may influence both human discourse and model behaviour [26].

Accordingly, the results of this study should be interpreted as a proof of concept for the generation of post-reflection dialogue within a constrained experimental setting. Further work is required to evaluate the robustness of the approach across broader datasets, additional languages, and alternative institutional contexts, as well as to explore issues of calibration, value alignment, and the possibility of cross-context generalisability.

This limitation aligns with prior work suggesting that no universal solution currently exists for all applications of artificial intelligence systems, particularly in socially situated or value-sensitive domains [18, 24]. Similarly, it remains an open question whether a universal approach to reflective dialogue generation is achievable across heterogeneous contexts, or whether performance necessarily depends on domain-specific adaptation.

From a design perspective, this raises a trade-off between generalisability and contextual performance in systems intended for deliberative or reflective tasks. Addressing this trade-off requires careful consideration of whether system design should prioritise broad applicability across domains or optimised performance within specific settings. This tension reflects a broader characteristic of communicative and socio-technical systems rather than a limitation specific to the present study.

5.5 Towards the Future

This study was not exhaustive, as mentioned in the aforementioned subsections, and as such future work can definitely be done in order to improve upon the concepts explored here. For example, different prompting variations within the single-turn and multi-turn approaches could be analysed as well as different transcript contexts to pinpoint whether the overall increase in perceived satisfactions was attributed to the specific combination of data, models, and prompts used within this study, or whether it is intrinsic to LLM architecture and the nature of conversation. Future work could also be conducted to determine to what extent the language present within the speech or transcripts analysed matters in relation to the perceived effectiveness of the prompting or reflection strategy. Due to Large Language Model tokenisation schemes usually being aimed at the English alphabet and common patterns of elicitation present within the lexicon, it is possible that the results observed within this study can skew to be more favorable or less favorable depending on the language[13, 31] and the answer might not be as simple as making the model accommodate more languages[9]. At the same time, as has been espoused throughout this paper: just as it is unfeasible to consider

only monolithic solutions that can be deployed indiscriminately to receive optimal answers, it is likewise unfeasible to consider developing only solutions that are able to function properly in all languages for the purposes of deliberation and self-reflection[38]. To that end, work should not be neglected in either aspect: both integrated solutions and specialised solutions should be investigated in order to not tunnel vision on a single expected answer.

Just as language and context are important in this study, so were the Large Language Models used. The models were selected because they represent distinct open-weight LLM ecosystems while maintaining comparable levels of capability[3, 11, 12]. The use of multiple model families reduced the risk that observed effectiveness of reflection are merely artifacts of a single architecture or prompting procedure. Rather than evaluating the absolute performance of any one model, the objective of this research was to investigate whether the proposed reflective dialogue methodology exhibits robustness across models with differing design philosophies and reasoning styles, within a Deliberative AI approach. It is, however, undeniable that the models used were not exhaustive. Further research can, as such, also be done in order to investigate whether a link can be found between the models used and the perceived effectiveness of the reflective dialogue evaluations, independent of the prompting strategy used.

6 Conclusion

In this paper, the possibility of LLM-based tools generating reflective dialogue that can enhance the interlocutor's perceived satisfaction in relation to a set of values and points of contention was explored with two distinct approaches of single-turn dialogue and multi-turn dialogue. The approaches were explored via the prompting of dialogue initiators and Synthetic Personae evaluators across several transcripts from a homogenous context, with parameters ensuring the reproducibility as well as variance of the results. The resulting evaluations were subsequently aggregated across several dimensions and analyzed in order to see the relation between the two approaches as well as the perceived satisfaction for each value in turn. Taken together, the two sub-questions explored address both whether post-reflection dialogue generation is achievable and how it can be realised in practice, thereby providing an answer to the main research question. We were able to find that perceived satisfaction generally improved with more iterations and interactions between the interlocutor and system. With the analyses presented within the previous sections as well as the recommendations for future work, we believe this research can serve as a useful foundation for future work on Deliberative AI applications and development, as well as prove informative for subsequent studies on deliberation in relation to Large Language Models as a whole.

7 Responsible Research

7.1 Data Processing and Potential Risks in Deployment

The ethical aspects of this research do not indicate any direct harm-mental or otherwise-arising from the execution of the research. The transcripts used for the analysis of modern deliberation and for the generation of reflective dialogues were publicly available and anonymised, thereby reducing the risk of harm to identifiable individuals based on the experiments conducted. In addition, all LLM-based systems were executed locally on a secured

device without network connectivity, which eliminated risks related to external data transmission, logging, or third-party retention of inputs.

However, these constraints do not eliminate ethical considerations for downstream applications of the proposed methodology. In particular, if similar reflective dialogue systems were deployed in real-world settings such as educational feedback systems, political deliberation support tools, or organisational decision-making assistants, several risks identified in prior literature may become relevant. These include overreliance on model-generated reasoning in high-stakes contexts, where users may defer to outputs despite known issues of hallucination, uncertainty calibration, and still existent loss of agency [6, 8, 10, 14], value misalignment in evaluative settings, where LLM outputs reflect training-data biases rather than stakeholder values [5], and context sensitivity and prompt dependency, where small changes in framing can lead to materially different outcomes in reasoning or evaluation despite the apparent same problem specification being given[30, 37].

In applied deployments, these limitations could manifest differently depending on the domain. For example, in civic deliberation tools, biased or overly confident summaries of stakeholder positions could distort perceived consensus. In educational contexts, reflective feedback generated by such systems may inadvertently steer learners toward predefined interpretations rather than supporting open-ended reasoning. These risks suggest that the methodology should be considered suitable primarily for exploratory or assistive use cases rather than as a substitute for human evaluators in consequential decision-making settings.

7.2 Availability of Large Language Models Used for Reproducibility

The three models that were used during this research, as previously mentioned, were the Llama 3.2 3B, Gemma 4 E2B, and Qwen 3 V1 4B Models. These models are open source, and were run locally via the software LM Studio. The models themselves occupy 1.88, 5.89, and 3.3 gigabytes respectively. Several other open source solutions exists that can be used to run these models locally in the case of LM Studio's unavailability such as Ollama and GitHub, GitLab repositories. LM Studio allows for direct modification of an LLM's operational parameters such as temperature, within the software but in order to modify them on Ollama, for instance, terminal commands are needed. The exact parameters used were as such: each pair of LLMS operated on the same seed range of 42-51 on each transcript sequentially, with a temperature of 0.3, Top K sampling of 40, Repeat Penalty of 1.1, Top P Sampling of 0.95, and Min P Sampling of 0.05. All of these models were run on a device sporting 16 gigabytes of RAM.

7.3 Use of Synthetic Personae

Another ethical consideration concerns the extent to which findings derived from Synthetic Personae can be generalised to human evaluators. Synthetic Personae were found Synthetic Personae were used as proxies for human interlocutors in this study due to the absence of human participant data, and were instantiated to approximate interlocutor perspectives within the constraints of the available transcript information. The "Real Conversations" configuration was selected because it constructs personae from observed conversational data rather than explicitly specified demographic or psychological attributes, thereby grounding the evaluation in the content and context of the interaction.

The deployment of Synthetic Personae followed the Persona Transparency Checklist

framework, and design choices described in the Method section were applied to ensure consistency across experimental conditions. Each transcript used in the study contained sufficient contextual information to support persona instantiation while remaining bounded to reduce dilution effects from extended interaction histories.

It is acknowledged that Synthetic Personae do not constitute direct representations of human cognition or of their thinking patterns, and therefore cannot fully capture the complexity of human evaluative behaviour. As a result, conclusions drawn from these evaluations should be interpreted as model-based approximations rather than direct measurements of human response. This constraint is inherent to the use of simulated evaluators in place of human participants.

Repository and Reproducibility

The dataset used in this study is publicly available[1], alongside the prompts and experimental outputs generated during the research. The methodological procedures employed to obtain these results are described in detail within the main body of this paper, enabling partial replication of the experimental pipeline.

It is acknowledged that certain external dependencies, including specific large language models or systems used for reflective dialogue generation and Synthetic Persona evaluation, may become unavailable or evolve over time. This reflects a general limitation in computational research involving third-party or rapidly developing models, where reproducibility may be affected by changes in model availability or behaviour.

To mitigate this issue, all materials required to reproduce the experiments have been documented and made available where possible. This includes the anonymised transcripts, prompt formulations, and the codebook used for evaluation. These resources are intended to support replication and extension of the presented methodology, subject to the availability of the underlying model infrastructure.

All associated materials are accessible via the accompanying GitHub repository.[2]

7.4 Use of Generative AI

Generative Artificial Intelligence tools were employed exclusively for the execution of the experiments described in this research, namely the generation and evaluation of reflective dialogues using Large Language Models. Such systems were not used for the drafting of the paper, the conduct of the literature review, the formulation of research questions, or the derivation and interpretation of conclusions. Outside the experimental framework, their use was limited to minor auxiliary tasks, such as formatting bibliographic information into BibTeX entries for inclusion in the bibliography, and spell checking paragraphs of text. The analysis of results, interpretation of findings, and the written content presented in this thesis were not influenced by Generative AI.

8 Acknowledgements

This research was made possible with help from the team at TUDelft which comprises the supervision of Michael Grauwde and Responsible Professor Willem-Paul Brinkman as well as my colleagues within the TUDelft's CSE3000 Research Project course. I believe that our

[1] https://hansard.parliament.uk/Commons/2026-02

[2] https://github.com/claticiv/post-reflection-deliberative-ai-tudelft

work has been worthwhile, and that it will prove to be useful in further bridging the gap between Artificial Intelligence and humanity in a responsible, beneficial way.

I would like to thank my family and friends for their support during my education as well as their insights which have shaped this research at every step of the way.

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] J. Batzner, V. Stocker, B. Tang, A. Natarajan, Q. Chen, S. Schmid, and G. Kasneci. Whose personae? synthetic persona experiments in llm research and pathways to transparency. In *Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)*, 2025.

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[6] Zana Bucinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[7] Thomas K. Burch. Longitudinal research in social science: Some theoretical challenges. *Canadian Studies in Population*, 28(2):263–283, 2001.

[8] Irina Carnat. Human, all too human: accounting for automation bias in generative large language models. *International Data Privacy Law*, 14(4):299–314, 2024.

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[10] Nicole Cruz. Illusions of understanding from outsourcing thinking to llms. *Computational Brain & Behavior*, 2026.

[11] Abhimanyu Dubey et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[12] Gemma Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[13] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjan Krishna, Marc'Aurelio Ranzato, Francisco GuzmÃ¡n, and Angela Fan. *Few-shot Learning with Multilingual Language Models*, 2021. arXiv preprint arXiv:2112.10668.

[14] Lujain Ibrahim, Katherine M. Collins, Sunnie S. Y. Kim, et al. Measuring and mitigating overreliance is necessary for building human-compatible ai. *arXiv preprint arXiv:2509.08010*, 2025.

[15] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3 edition, 2026. Online manuscript released January 6, 2026.

[16] Rafal Kocielnik, Lu Xiao, Daniel Avrahami, and Gary Hsieh. Reflection companion: A conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–26, 2018.

[17] Rachel B. Larson. Controlling social desirability bias. *International Journal of Market Research*, 61(5):534–547, 2019.

[18] S. Ma, Q. Chen, X. Wang, C. Zheng, Z. Peng, M. Yin, and X. Ma. Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23, April 2025.

[19] Jane Mansbridge, James Bohman, Simone Chambers, David Estlund, Andreas Follesdal, Archon Fung, Cristina Lafont, Bernard Manin, and Jose Luis Marti. The place of self-interest and the role of power in deliberative democracy. *Journal of Political Philosophy*, 18(1):64–100, 2010.

[20] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[21] Michael R. Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi for operationalizing progress on the path to agi. *arXiv preprint arXiv:2311.02462*, 2023.

[22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

[23] Joon Sung Park, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

[24] Sangho Park, Hari Subramonyam, and Chinmay Kulkarni. Thinking assistants: Llm-based conversational assistants that help users think by asking rather than answering. *arXiv preprint arXiv:2312.06024*, 2023.

[25] Delroy L. Paulhus. Measurement and control of response bias. In John P. Robinson, Phillip R. Shaver, and Lawrence S. Wrightsman, editors, *Measures of Personality and Social Psychological Attitudes*, pages 17–59. Academic Press, San Diego, CA, 1991.

[26] Siddhesh Pawar, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. Presumed cultural identity: How names shape llm responses. *arXiv preprint arXiv:2502.11995*, 2025.

[27] Sandra Peter, Kai Riemer, and Jevin D. West. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22):e2415898122, 2025.

[28] Uwe Peters. Explainable ai lacks regulative reasons: Why ai and human decision-making are not equally opaque. *AI and Ethics*, 3:963–974, 2023.

[29] Robert C. Post. Three concepts of privacy. *Georgetown Law Journal*, 89(6):2087–2098, 2000.

[30] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv preprint arXiv:2102.07350*, 2021.

[31] Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3118–3135, 2021.

[32] P. S. Sachdeva and Tom van Nuenen. Deliberative dynamics and value alignment in llm debates. *arXiv preprint arXiv:2510.10002*, 2025.

[33] Alireza Salehi, Maryam Marzban, Maryam Sourosh, Farkhondeh Sharif, Mahmoud Nejabat, and Mohammad Hadi Imanieh. Social well-being and related factors in students of school of nursing and midwifery. *International Journal of Community Based Nursing and Midwifery*, 5(1):82–90, 2017.

[34] S. F. Sanchez. Llm-based simulations of human behavior in psychological research. In *Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)*, 2025.

[35] Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. Large language model routing with benchmark datasets. *arXiv preprint arXiv:2309.15789*, 2023.

[36] Bibi van den Berg, Pauline Hutten, and Ruth Prins. Security and safety: An integrative perspective. In Gabriele Jacobs, Ilona Suojanen, Kate E. Horton, and Petra Saskia Bayerl, editors, *International Security Management: New Solutions to Complexity*, pages 13–27. Springer International Publishing, Cham, 2020.

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[38] David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[39] L. Xi, Y. Zhang, and Q. Wang. Investigating the effects of an llm-based socratic conversational agent on students' academic performance and reflective thinking in higher education. *Computers and Education*, page 105494, 2025.

A Appendix

Comprises the tables holding the average evaluation metrics across all eight transcripts for the single-turn approach and multi-turn approaches, respectively.

Table 3: Single-Turn Aggregated Results across all Transcripts across all five metrics

Gen / Eval	**Safe**	**Priv**	**Auto**	**Soc. Well-being**	**Cont.**	**Avg.**
Llama 3.2 → Gemma 4	8.0250	6.1000	7.1500	8.4875	8.6000	7.6725
Llama 3.2 → Qwen 3 V1 4B	8.6375	7.5250	8.2125	9.0125	7.4875	8.1750
Gemma 4 → Llama 3.2	7.5000	5.8125	7.3750	7.6875	6.8625	7.0475
Gemma 4 → Qwen 3 V1 4B	8.3750	6.1250	7.8750	8.7500	7.1250	7.6500
Qwen 3 V1 4B → Llama 3.2	6.1125	4.1375	5.1375	5.4875	7.8500	5.7450
Qwen 3 V1 4B → Gemma 4	7.7625	5.6500	6.4875	7.7250	9.1125	7.3475
Column Average	**7.7354**	**5.8916**	**7.0395**	**7.8583**	**7.8395**	**7.2729**

Table 4: Multi-Turn Aggregated Results across all Transcripts across all five metrics

Gen / Eval	**Safe**	**Priv**	**Auto**	**Soc. Well-being**	**Cont.**	**Avg.**
Llama 3.2 → Gemma 4	8.1000	6.6500	7.7625	8.3375	9.3625	8.04305
Llama 3.2 → Qwen 3 V1 4B	7.2375	5.4000	6.7625	7.8500	6.9250	6.8350
Gemma 4 → Llama 3.2	7.8750	7.2500	7.9750	8.8750	8.3750	8.0700
Gemma 4 → Qwen 3 V1 4B	8.9875	8.0625	9.3750	9.4875	7.2625	8.6350
Qwen 3 V1 4B → Llama 3.2	7.3625	7.3250	7.3875	8.3625	8.7250	7.8325
Qwen 3 V1 4B → Gemma 4	8.7375	7.6000	8.4875	9.0250	9.7250	8.7130
Column Average	**8.0500**	**7.0479**	**7.9593**	**8.6462**	**8.3958**	**8.02142**