How to design post-reflection dialogue from transcripts using identified values, value tensions, and consensus points with a deliberative approach in public safety scenarios?

Supervisor Michaël Grauwde
Responsible Prof. Dr. W. Brinkman

Victor-Emanuel Clatici, TUDelft CSE BSc

OVERVIEW

Introduction:

- Most contemporary LLMs and LLM-based tools are aimed at providing the user with "fixed recommendations they can accept or reject".[1] This design choice is not always ideal in social or deliberative scenarios where interaction with the user is more encouraged.
- This research explored this exact need of innovation by using Deliberative AI techniques with two approaches to answer the research question, assessing the feasibility of LLM-generated reflective dialogue while analyzing differences in evaluation between various dialogue creation methods.

Method:

- Transcripts were selected from a homogenous deliberative context, anonymized and given to both the Generator and Evaluator for each experiment.
- For each of the transcripts the seed range was varied across identical temperatures and model parameters to ensure variance within the results as well as reproducibility.
- The single-turn and multi-turn approaches simulated short and longer conversations respectively with the aim of uncovering potential differences in the effectiveness of deliberation.

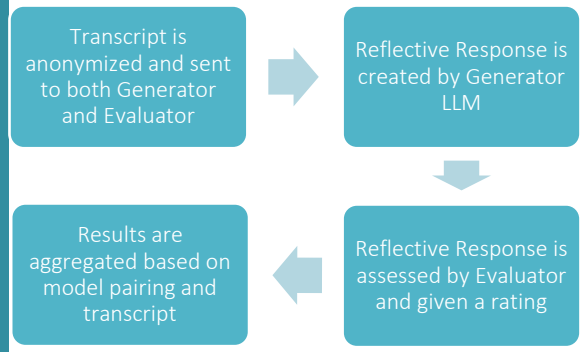
Evaluation:

- The Synthetic Personae constructed based on interlocutors from the transcripts assessed the reflective response or discussion respectively and gave ratings based on the four values of safety, privacy, autonomy, societal well-being, as well as the fifth criterion of points of contention.

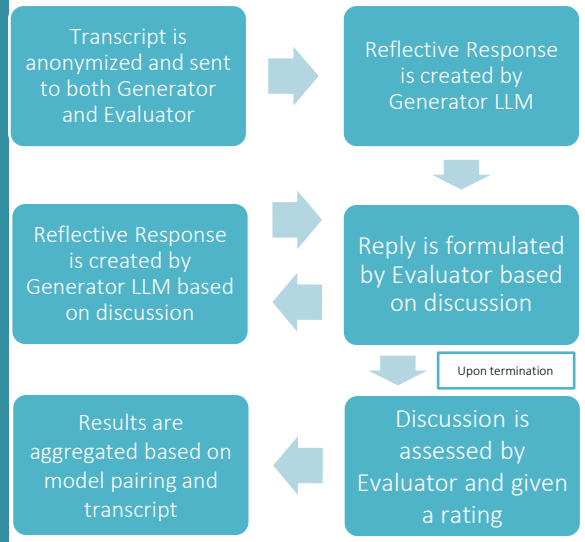
[1] Ma et al. (2025). *Towards Human-AI Deliberation*. CHI '25. https://doi.org/10.1145/3706598.3713423

PROCESS

Single-turn approach



Multi-turn approach



The evaluations were on a scale from 1 to 10 for each of the four reviewed values and for the fifth criterion of points of contention.

RESULTS

Single-turn approach

Gen/Eval	Llama 3.2	Gemma 4	Qwen 3 V1
Llama 3.2	N/A	7.6725	8.1750
Gemma 4	7.0475	N/A	7.6500
Qwen 3 V1	5.7450	7.3475	N/A
Average	**7.2729**		

Multi-turn approach

Gen/Eval	Llama 3.2	Gemma 4	Qwen 3 V1
Llama 3.2	N/A	8.0430	6.8350
Gemma 4	8.0700	N/A	8.6350
Qwen 3 V1	7.8325	8.7130	N/A
Average	**8.0214**		

Aggregated results per criterion

Method	Safety	Privacy	Autonomy	Soc.	Cont.
Single Turn	7.7354	5.8916	7.0395	7.8583	7.8395
Multi Turn	8.0500	7.0479	7.9593	8.6462	8.3958
Average	**7.8927**	**6.4697**	**7.4994**	**8.2522**	**8.1176**

The results are favorable with respect to the research question, providing evidence that the chosen approaches enable meaningful deliberation.

TAKEAWAYS AND LIMITATIONS

Conclusion:

- Effective reflective dialogue creation was found to be possible, with the produced dialogue having average ratings in both approaches exceeding the predefined success threshold of 5/10.
- Every model combination in both approaches barring a single value had an average rating for each criterion exceeding the threshold of 5/10, suggesting that the deliberative method proved to be capable of creating useful reflective dialogue.
- The ratings on average and aggregated across all model combinations were higher for the multi-turn approach, indicating the possibility that longer conversations that engage with both parties for longer periods yield more meaningful deliberation.

Limitations:

- The extent to which the English language, the cultural and socio-political context of the transcripts, and the prevalence of each value in the initial transcripts influence deliberation efficiency has yet to be determined.
- As the study was conducted using a non-exhaustive set of open-source LLMs, the extent to which the selection and ordering of the models causally influence the perceived effectiveness of the dialogue has yet to be established.