Pillars	***Description***	***Checklist Category***	***Checklist***	***Result***
Consistency	The findings of human evaluation must be reliable and generalizable.	Ill-defined or complex evaluation guidelines	- Mechanism for reporting issues & ambiguities	'Reviewer Notes' section is provided for each data point for evaluators to submit a text form entry.
			- Expert annotators independently able to follow annotation guidelines and achieve higher inter-annotator agreement (IRA)	Evaluators were able to independently follow the guidelines, however high IRA was not achieved.
		High task complexity	- Each single task consists of only one question	Each single task is focused on only one aspect of the data.
			- Evaluation tasks broken down into easier tasks	Evaluation tasks are simplified to be in the form of questionnaire-like with Likert-scales, in which each scale contains concise 1 sentence description.
		Ill-suited evaluators	- Verified qualification of evaluators	Evaluators are workers within the company with domain knowledge and experience, and are currently using and implementing the knowledge within the domain in their daily work tasks.
			- Exam/attention check quality control metrics during evaluation	Due to the size of data that is considerably small, due to time and resources constraints, this is not implemented, however future evaluations can make use of the

				addition of this metric.
		Small number of evaluators and/or test set	- Justifiable number of evaluators	The number of evaluators recruited are 2 evaluators, where the evaluation may be less robust, therefore, increasing the number of evaluators can be implemented to test whether agreement remains stable across more raters.
			- Justifiable size of dataset	The number of data points evaluated were 25, where a quick overview of the chatbot's performance on the included scenarios can be gained, however the results cannot be representative of the chatbot's capability as a whole.
		Rating scales such as the Likert	- Usage of proper scoring method to measure perception and non-perception aspects	Likert scales were used to assess perception-based aspects of the chatbot, and categorical labels were used to assess judgement-based aspects.
		Inter-rater agreement	- Usage of IRA measure	The IRA measures used are Percent Agreement, Cohen's Kappa, and Pearson Correlation.
			- Analysis on item-wise IRA	Considering Cohen's Kappa scores, all criteria have very low scores. However, with percent agreement it is shown that criteria: tone and style appropriateness, relevance, and clarity have relatively lower percentages, as the

				aspects are more perception-based (opinion sensitive). However, as overall scores are still considerably low, standardization is the main point to improve from the evaluation design.
			- Standardized qualifications of evaluator	Recruited evaluators have roughly the same qualifications in relation to the domain knowledge, therefore increasing IRA would require further standardization of the rubric.
			- Continuous measure of IRA for each evaluation	Currently, the human evaluation has only been done once, however it would be recommended to do more rounds in future development processes, where IRA should be measured each time until higher IRA is achieved, proving the rubric is sufficiently standardized.
Scoring Criteria	The scoring criteria must include both general purpose criteria such as readability, as well as be tailored to fit the goal of the target tasks or domains.		- Evaluation includes typical dimensions: Fluency, Coherence, Relevance, Factuality	The evaluation includes the mentioned dimensions represented by factual correctness, tone and style appropriateness, clarity, and relevance scores present in the rubric.
			- Presence of multi-dimensional domain-specific criteria	Evaluation of each response was done on multiple criterias including: factual correctness, tone and style appropriateness, safety score, relevance, and

				clarity.
			- Responsible AI criteria evaluation	A responsible AI criteria was included, labeled as safety score, where the criteria assess if the response may produce harmful activity when put into practice.
Differentiating	The evaluation test sets must be able to differentiate the various capabilities as well as the weaknesses of generative LLMs		- Evaluated criteria and test cases used	For each criteria, 25 test cases were used.
			- Test on end user use cases	The end user use cases that were included were cases where informational response (containing technical knowledge as per request) was generated. As there is presence of time and resources constraints, the evaluation was focused on this use case, as the reliability of the information generation is crucial to assess. Other use cases are more towards usability cases.
			- Consideration on possible test cases having been used in training data	As the model used is OpenAI's GPT model, where it was trained with billions of data from the internet, there may be a possibility that the test cases have been used to train the model, thus the inconsistent performance of the chatbot in interpreting the knowledge context it retrieved.
			- Implementation of robustness tests to evaluate model weaknesses	Due to time and resources constraints, robustness tests were not

				implemented yet.
User Experience	The evaluation must take into account user experience, including their emotions & cognitive biases, when designing experiments and interpreting results.		- Implementation of denoising algorithms applied on the rating-based metrics	The Likert scale was used to account for human cognitive uncertainty in perception-based metrics.
			- Splitting of content for factual check and truthfulness evaluation	This was not implemented, as for now it is still considered not necessary due to the nature of the evaluators that have the domain knowledge already.
			- Proper use of the Likert-scale for the correct measure	Likert-scale was used correctly for measuring perception-based metrics (more opinion sensitive).
			- Test of actual usability of the output	The actual usability of the output was not tested, as adding this instruction into the evaluation would mean raising the time needed to complete the evaluation process, whereas time and resources constraints were already present.
Responsible AI	The evaluation needs to account for responsible AI including aspects such as bias, safety, robustness, and privacy capabilities of the model.		- Performance of safety testing	Safety testing was implemented by including the safety score metrics for the evaluators to assess the responses on.
			- Performance of privacy testing	Privacy testing was not implemented, however it is for now not deemed necessary as it is unlikely that there are scenarios where users are required to leak sensitive information to the chatbot.

			- Testing on bias	Testing on bias was not implemented, however in the future development process, it may be useful to test, to have the chatbot be more accessible to users who may have disabilities.
			- Number of evaluators and the diversity	In total, two evaluators were recruited, both having roughly similar qualifications in terms of domain knowledge. However, the limited age and gender diversity among the evaluators may reduce the external validity and generalizability of the evaluation results.
Scalability	Human evaluation must be scalable for pragmatic widespread adoption.		- Optimization of usability aspects for the annotators to reduce annotation time	The evaluation process utilized Langsmith as the interface to allow for a simpler and faster process, as the interface provides the data and rubric side by side.
			- Presence of automation in the human evaluation	Currently, no automation was implemented in the evaluation process.