Instructions:

- Create prompt
    - Ensure the responses correspond to a chatbot and not a human
- Create test set
    - Ensure gold response aligns with a bot
    - Add few-shot examples for each context-response pair in the test set
    - **If the agent utterance mentions any points that might suggest the agent is human, remove them from consideration**
    - Create samples of varying response lengths (10 - 15 - 20)
- Create 15 test samples per test set
    - Four context response pairs
    - All context-response pairs correspond to similar generated response type
    - Zeroshot - 1 context - create a response
    - Oneshot = 1context-response pair as an example - 1 context - generate a response
    - Twoshot = 2context-response pair as an example - 1 context - generate a response
    - Threeshot = 3context-response pair as an example - 1 context - generate a response
- Test for all the above-stated models.
- Fill in all the configuration details in a combined spreadsheet.

Evaluation:

- Tokens-based- BLEU, METEOR, etc
- Embedding-based- Vector Extrema, BERT score, etc
- Human evaluation- fluency, completeness, etc
- Comparison in between models wrt. Different aspects
- Custom evaluation criteria (context relevance, length, etc., with their combinations)
- Pros and cons corresponding to all models

Prompt generation	"System prompt based on the task asssociated with the dataset, Gold Response based on intent, Context based on the gold response (5 - 7 previous utterances) Knowledge based on the product description (dataset specific) Few shot samples based on the same intent flow as the context and gold response

	All are combined to generate the prompt: (few shot samples depends on test config) For example: Selection criteria for context: intent == 'counter-price' AND seller == 'buyer' AND response_intent == 'counter-price':"
Evaluation Criteria	Test set of 100 dialogues 0-shot, 1-shot and 2-shot Prompts Models: LLama2, Falcon Dataset: MHLCD Automatic Evaluation Metrics Generic Metrics BLEU, PPL, etc. Response Length: no. of tokens in the generated utterance Task-specific Metrics Pol: no. of polite utterances generated Emp: no. of empathetic utterances generated Human Evaluation Metrics Generic Metrics Fluency; depends on the respinse statement only; 1-5 i.e. Gibberish to proper response Fluency - 5: Flawless, 4: Good, 3: Non-native, 2: Disfluent, 1:Incomprehensible; Adequecy; wrt reference; are all the slots/information in the reference present in the generated response? (1-5) Adequacy - 5: All, 4: Most, 3: Much, 2: Little, 1: None Correctness; wrt context; are all the slots/information in the asked in Query and context present in the generated response? (1-5) Correctness - 5: All, 4: Most, 3: Much, 2: Little, 1: None

	Task-specific Metrics: Politeness Empathy
Auomatic metrics	BLEU, Greedy-Avg-vectorextrema, BERT Score
	LEN, Emotion Probability
Negotiation	
	Negotiation Strategies: (i). Negotiation Consistency (N- Con): It is the measure of consistency (absence of arbitrariness) in the negotiation approach within a dialogue (ii). Bargaining Efficacy (B-Eff): It measures the ability of the negotiation system to present compelling arguments, reasoning, or incentives that influence the other party's decision-making process., (iii). Outcome fairness (O-fair): It assesses the fairness or equity of the final outcomes reached during the neg

	(iv). Dialogue- fluency (D-F): It measures the overall grammatical correctness of the generated responses, and (v). Dialogue-Engagingness (D-E): Measures the extent to which a conversation or dialogue is interesting, captivating, and able to hold the attention of the participants. The evaluators assigned scores on a scale of 1 to 3 for each metric (The higher the better).