Model	Session 2			Session 3			Session 4			Session 5		
	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L
BlenderBot	4.71	1.47	18.20	3.85	0.93	17.10	3.69	0.83	16.78	4.00	1.19	17.19
$\text{BlenderBot}_{\text{msc}}$	6.39	2.56	19.30	5.82	1.93	18.67	5.30	1.76	17.9	6.10	2.30	18.65
FID-RAG	6.41	2.51	19.82	5.83	1.95	18.38	**5.81**	1.85	**18.44**	6.02	2.27	18.52
HAHT (ours)	**6.69**	**2.73**	**20.02**	**6.03**	**2.20**	**18.70**	5.48	**1.95**	18.00	**6.38**	**2.51**	**19.18**

Table 4: Automatic evaluation results of different models on session-opening data. Session i indicates there are i-1 history conversation sessions. B-2, B-3, and R-L denote BLEU-2, BLEU-3, and Rouge-L respectively. The best results are in **boldface**.

Model	Session 2			Session 3			Session 4			Session 5		
	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L
HAHT	**5.07**	**1.57**	**16.90**	**5.27**	**1.67**	**16.72**	**5.00**	**1.55**	**15.97**	**5.16**	**1.60**	**16.42**
$\text{HAHT}_{\text{w/o HIER}}$	5.00	1.57	16.72	5.19	1.63	16.61	4.86	1.49	15.90	5.10	1.57	16.21
$\text{HAHT}_{\text{w/o HIST}}$	4.98	1.50	16.81	5.09	1.58	16.51	4.75	1.45	15.51	5.10	1.49	16.24
$\text{HAHT}_{\text{w/o SW}}$	5.01	1.56	16.86	5.19	1.61	16.46	4.87	1.55	15.88	5.07	1.55	16.17

Table 5: The performance achieved by HAHT and different HAHT variants. Session i indicates there are i-1 history conversation sessions. B-2, B-3, and R-L denote BLEU-2, BLEU-3, and Rouge-L respectively. The best results are in **boldface**.

5.2 Human Evaluation

Table 3 summarizes the human evaluation results on the Facebook MSC dataset. Generally, HAHT outperforms all the baseline methods in terms of all perspectives. This observation is consistent with the automatic evaluation results shown in Table 2. In particular, we find that HAHT performs much better than other baselines in terms of history relevancy. This demonstrates that HAHT can better leverage the history conversation sessions and engage the user more in the on-going session with the history memory. HAHT also performs better than other baselines in terms of readability and context relevancy. This indicates that HAHT can better understand the current conversation context with the help of the history memory.

5.3 Evaluation on Session Openings

In the MSC task, the session opening is the first conversation turn of the current conversation. According to our observation and the similar observation in (Xu et al., 2022), the opening conversation turn is categorically different from other conversation turns. It typically involves a statement or question that aims to reengage the other speaker based on the known information that has been exchanged in history conversations. Therefore, the performance on the session opening data can further demonstrate the model's capability in understanding and leveraging history conversations.

We compare all models on these opening responses and show the results in Table 4. We observe that the proposed HAHT model achieves the best performance in terms of most metrics. Especially, when there are 4 history conversations, HAHT outperforms FID-RAG and $\text{BlenderBot}_{\text{msc}}$ by 10.6% and 9.1% in terms of BLUE-3. This indicates that the proposed HAHT can better leverage conversation history to reengage the user into a new conversation session.

5.4 Ablation study

To better understand the effectiveness of each main component of HAHT, we conduct ablation study for HAHT. Specifically, we consider the following variants of HAHT.

- **$\text{HAHT}_{\text{w/o HIER}}$**: In this variant, we do not encode the history conversations hierarchically. Instead, we concatenate all the utterances of history conversations into a long sentence and directly encode it using the transformer encoder.
- **$\text{HAHT}_{\text{w/o HIST}}$**: In this variant, we remove the history encoder from HAHT.
- **$\text{HAHT}_{\text{w/o SW}}$**: In this variant, we remove the switching mechanism from the response generator of HAHT.

Table 5 summarizes the results achieved by different HAHT variants, in terms of BLEU-2, BLEU-3, and Rouge-L. We note that HAHT outperforms $\text{HAHT}_{\text{w/o HIER}}$, which indicates that hierarchically