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	2.79	0.65	13.73	2.41	0.45	13.06	2.14	0.39	12.76	2.26	0.45	12.75
$\text{BlenderBot}_{\text{msc}}$	4.76	1.51	16.18	5.03	1.61	16.39	4.78	1.49	15.56	4.98	1.48	16.10
FID-RAG	4.82	1.54	16.53	5.04	1.61	16.42	4.84	1.48	15.89	5.06	1.57	16.01
HAHT (ours)	**5.07**	**1.57**	**16.90**	**5.27**	**1.67**	**16.72**	**5.00**	**1.55**	**15.97**	**5.16**	**1.60**	**16.42**

Table 2: Automatic evaluation results of different models on all session 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	Readability	Context Relevancy	History Relevancy
BlenderBot	1.78	1.13	0.09
$\text{BlenderBot}_{\text{msc}}$	1.82	1.56	0.13
RAG-FID	1.89	1.84	0.21
HAHT (ours)	**2.05**	**2.03**	**0.33**

Table 3: Human evaluation of the response generation by different methods. All scores are rated in four levels 0/1/2/3. The best results are in **boldface**. We measure the inter-rater reliability with Fleiss' Kappa (Fleiss and Cohen, 1973). Our annotations obtain "good agreement" for Readability (0.614) and "moderate agreement" for Context Relevancy (0.526) and History Relevancy (0.573).

versations. Only responses that are consistent with history conversations are considered relevant to history.

Each aspect is rated in four different levels 0/1/2/3, and the final score of each aspect is the average of the scores given by all annotators. We measure the inter-annotator reliability with Fleiss' Kappa (Fleiss and Cohen, 1973). For all evaluation metrics, the higher value indicates better performance.

4.3 Baseline Methods

We compare the proposed HAHT model with the following baseline methods.

- **BlenderBot** (Roller et al., 2021): This is a large-scale open-domain dialogue model pre-trained on the dialogue data scraped from social discussions on the web.
- **$\text{BlenderBot}_{\text{msc}}$**: This is the BlenderBot model finetuned on the MSC dataset.
- **FID-RAG** (Shuster et al., 2021): In this method, RAG-trained retriever (Lewis et al., 2020) is used to retrieve top-N history conversations, and Fusion-Decoder (FiD) (Izacard and Grave, 2021) is adopted to generate a final response considering the retrieved history conversations and current conversations. Following (Xu et al., 2022), N is empirically set to 5.

4.4 Model Settings

In this work, all the evaluated methods are trained following the same settings. Due to the limitation of computation resources, we use the BlenderBot model with 90M parameters as the initial pre-trained model and finetune it on the Facebook MSC dataset. The input length truncation is set to 256. The number of Transformer encoder layers n_{enc} and decoder layers n_{dec} are both set to 12. For model training, we use the Adamax optimizer (Kingma and Ba, 2014) with a learning rate of 1×10^{-6}, batch size of 16, dropout ratio of 0.1, and early stopping patience of 10. All the fine-tuned models are trained with a maximum of two 32GB GPUs (NVIDIA V100).

5 Experimental Results

This section presents the experimental results of the automatic evaluation, human evaluation, evaluation on session openings, ablation study, and case study.

5.1 Automatic Evaluation

The automatic evaluation results of different models are shown in Table 2. It can be observed that $\text{BlenderBot}_{\text{msc}}$ performs much better when fine-tuned on the MSC dataset. FID-RAG performs better than $\text{BlenderBot}_{\text{msc}}$. The potential reason is that RAG can retrieve important history conversations, and FID can combine the retrieved conversations with current conversations to generate better responses. Moreover, the proposed HAHT model consistently outperforms baseline methods in terms of all the evaluation metrics. This indicates that HAHT can better encode the history conversations, leverage history conversations to understand the current conversation context and generate more human-like responses.