encoding the history conversations can help the model reserve more history memory to generate more human-like responses. Moreover, HAHT achieves better performance than $\text{HAHT}_{\text{HIST}}$. This observation indicates that removing the history encoder causes the most decline in all metrics. This result confirms the necessity to leverage history conversations to understand the current conversation and generate the response. In addition, the performance degradation caused by removing the switching mechanism shows that directly copying words from the history conversation can help the model generate more history-aware responses.

5.5 Case Study

Table 6 shows a case study of the multi-session conversations generated by different models. Compared to baseline models, the proposed HAHT model can better leverage history conversations to understand the current conversation context and generate more history-aware responses. When the user discusses preparing sandwiches and lemonade with the agent, "*I can make sandwiches for us! I also have a very good recipe for homemade lemonade! Do you like lemonade?*", HAHT can remember information mentioned in the history conversations, such as the user likes reading and outdoor activities and it has adopted a book-lover persona before. HAHT can leverage these historical contexts and generate more human-like, context-relevant, and history-aware responses: *"I love lemonade! I'm sure we can find a lot of good recipes for sandwiches too. Sandwiches and lemonade are perfect for going outdoors and reading books."*.

6 Conclusion

In this work, we propose the History-Aware Hierarchical Transformer (HAHT) model for multi-session open-domain dialogue systems. The proposed HAHT model maintains a history memory by hierarchically encoding the history conversation sessions. After that, HAHT uses attention-based encoding to encode the current conversation context together with the history memory and updates the history memory with the current context. In order to explicitly leverage historical information in the responses, HAHT is designed with a history-aware response generator which can switch between a generic vocabulary and a history-aware vocabulary. Experimental results obtained under both normal and session opening MSC settings demonstrate that

History Conversation Sessions	
[Session 1]	
...	
User:	... My dad is a truck driver.
Agent:	... I am a **big reader** myself.
User:	... I like **working outdoors**.
...	
[Session 2]	
User:	What book are you currently reading?
Agent:	I'm reading a Stephen King novel. ...
...	
User:	I used to sit in the truck on **long trips** with my dad, and had nothing to do other than **read**.
Agent:	Ah, that makes sense. I started reading when my parents were at work. ...
Current Conversations	
...	
Agent:	... What kinds of picnic foods do you like?
User:	I can make sandwiches for us! I also have a very good recipe for homemade lemonade! Do you like lemonade?
Responses from Different Dialogue Systems	
BlenderBot:	I love lemonades ! I'm not sure if I could make a picnic for you though. I'd have to make a lot more than just a picnic.
$\text{BlenderBot}_{\text{msc}}$:	I love lemonade! I'm not a huge fan of lemonade, but I love lemonade.
FID-RAG:	I love lemonade! I'm sure we can find a good picnic for you.
HAHT (ours):	I love lemonade! I'm sure we can find a lot of good recipes for sandwiches too. Sandwiches and lemonade are perfect for **going outdoors** and **reading books**.

Table 6: A case study of an MSC with two history conversations. Only important utterances in the history and current conversations are presented. Complete conversations sessions are provided in Appendix A.1