Gaoling School of Artificial Intelligence
Renmin University of China

NEURAL INFORMATION PROCESSING SYSTEMS

StreamingDialogue: Prolonged Dialogue Learning via Long Context Compression with Minimal Losses

Jia-Nan Li* (lijianan@ruc.edu.cn), Quan Tu*, Cunli Mao, Zhengtao Yu, Ji-Rong Wen, Rui Yan

Jia-Nan Li
Renmin University of China

WeChat for Jia-Nan Li

Github Homepage for Jia-Nan Li

Background

- Traditional large language models fail to support lifelong conversations.
- Conversational LLMs with context comprehension, memory, and efficiency are gaining attention.
- Multi-turn dialogue LLMs like ChatGPT and MOSS are being released to wide acclaim.

Research on memory-capable streaming dialogue generation can lead to improved user experiences.

Challenges

- Fixed context length during pre-training restricts generation length.
- Standard attention complexity grows quadratically, increasing computational costs.
- Local attention leads to loss of dialogue memory and inconsistent context.

End-of-Utterance (EOU) often attracts more attention.

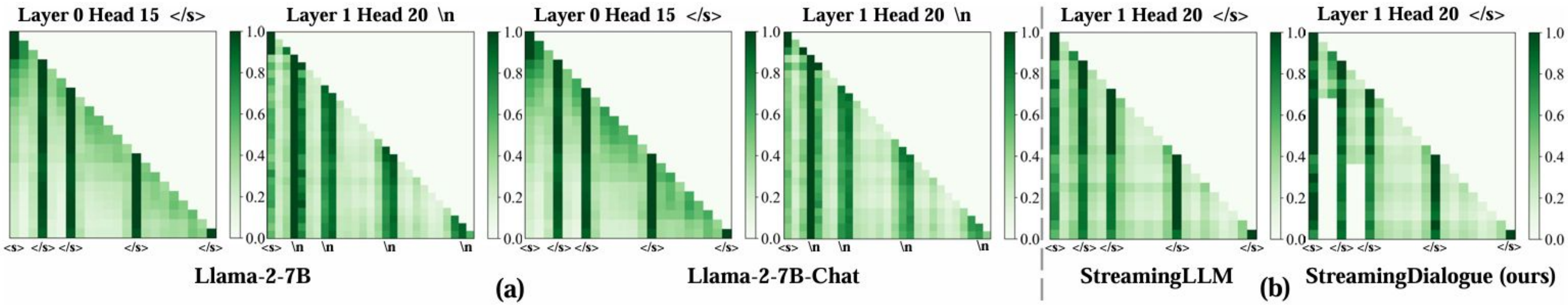


Figure 1: Attention map visualization. (a) Llama-2-7B/Chat with "</s>" and "\n" as EoU ("</s>" counts as one token, "\n" as two). (b) StreamingLLM versus StreamingDialogue attention on Llama-2-7B with "</s>" as EoU.

Method

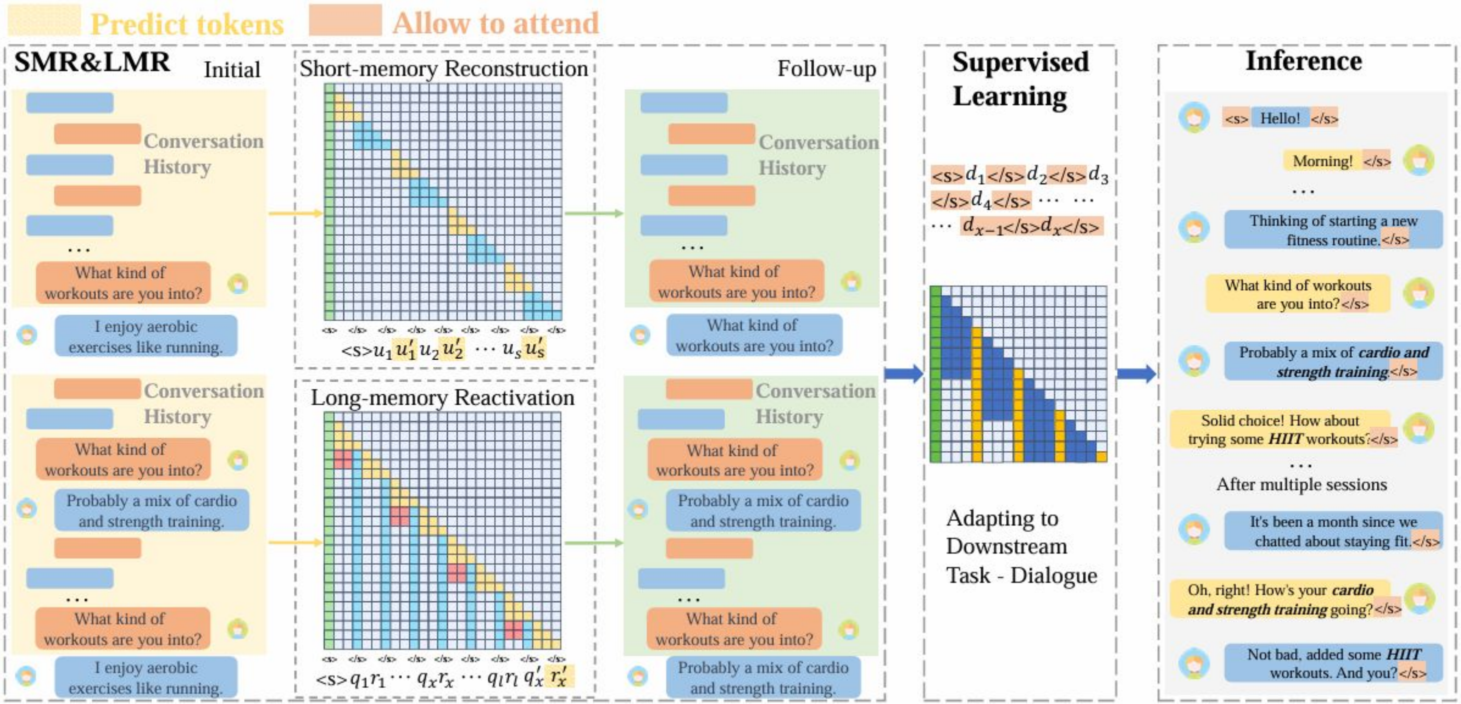
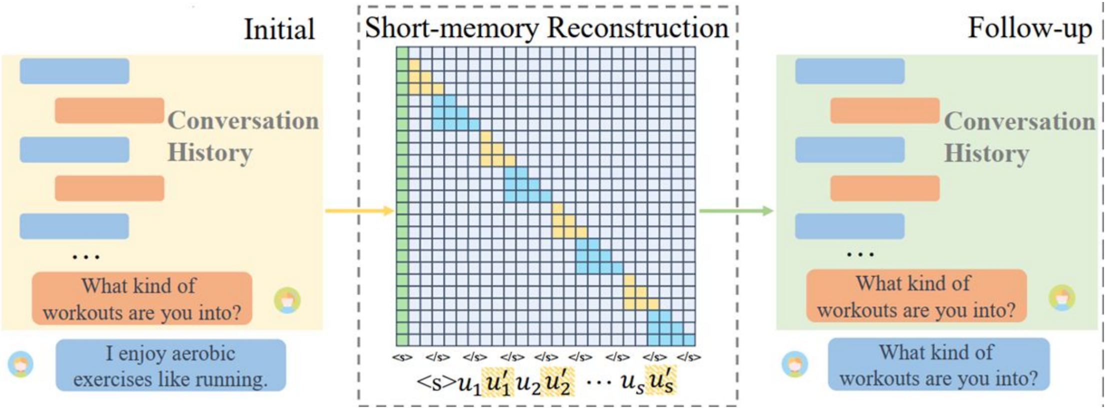


Figure 2: StreamingDialogue framework. SMR & LMR strategies co-train the model by adjusting attention mechanisms. In supervised learning, the SMR & LMR-trained model is fine-tuned with dialogue datasets. During inference, only specific tokens are cached, with critical historical dialogue information in bold italics for clarity.

- Short-memory Reconstruction
- Long-memory Reactivation
- Supervised Learning

Method

1. Short-memory Reconstruction



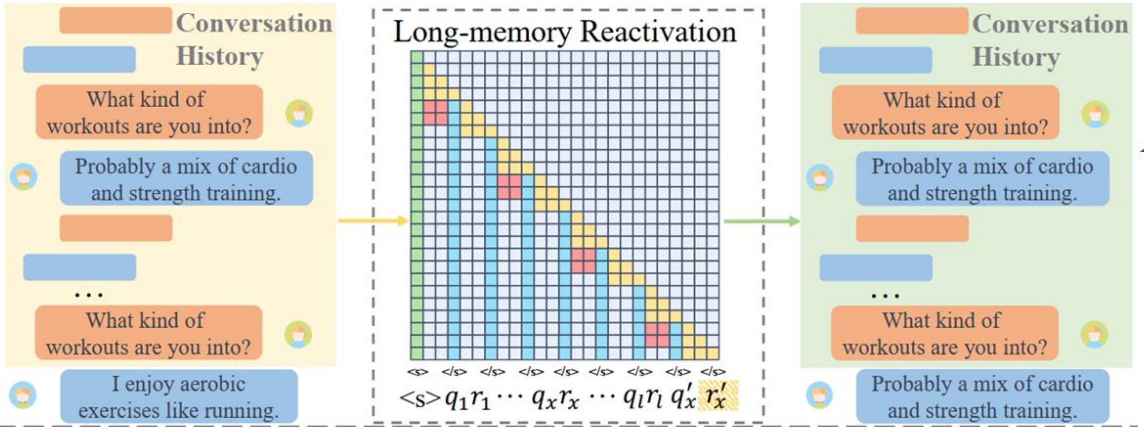
$$A_{ij} = \begin{cases} 1, & j = 0 \\ 1, & \lceil \frac{i}{l} \rceil = 2k \ (k \in \mathbb{N}^*), (\lceil \frac{i}{l} \rceil - 1) \cdot l \leq j \leq i < N \\ 1, & \lceil \frac{i}{l} \rceil = 2k + 1 \ (k \in \mathbb{N}), (\lceil \frac{i}{l} \rceil - 1) \cdot l < j \leq i < N \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathcal{L}_{\mathrm{SMR}} = - \sum_{(x,y) \in \mathcal{Z}} \sum_{t=1}^{|y|} \log \left(P_{\Phi} \left(y_t \mid x, y_{<t} \right) \right)$$

We design a dialogue reconstruction task aims at recreating dialogues based on EOUs after each sentence to enhance the EOUs' ability to aggregate information.

Method

2. Long-memory Reactivation



$$A_{ij} = \begin{cases} 1, & j = 0 \\ 1, & \lceil \frac{i}{l} \rceil = 2k \ (k \in \mathbb{N}^*), (\lceil \frac{i}{l} \rceil - 2) \cdot l < j \leq i < N \\ 1, & j = kl \ (k \in \mathbb{N}^*) \ \text{or} \ j > \lceil \frac{i}{l} \rceil - 1, \ j \leq i < N \\ 0, & \text{otherwise.} \end{cases}$$

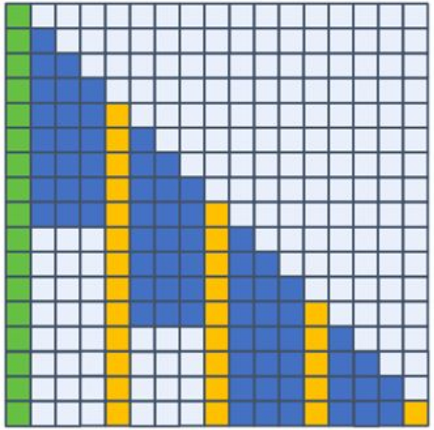
$$\mathcal{L}_{LMR} = - \sum_{(m,n)} \sum_{t=1}^{|n|} \log \left(P_\Phi \left(n_t \mid m, n_{<t} \right) \right)$$

We design a dialogue recall task that aims to recall and reproduce previously mentioned dialogues, enhancing the model's ability to extract historical information from EOUs and thereby improving long-term memory capabilities.

Method

3. Supervised Learning

<s>d_1</s>d_2</s>d_3
</s>d_4</s> $\cdots$ $\cdots$
$\cdots$ d_{x-1}</s>d_x</s>



$$A_{ij} = \begin{cases} 1, & j = kl \leq i \ (k \in \mathbb{N}),\ 0 \leq i < N \\ 1, & 1 \leq j \leq i \leq l \\ 1, & j \neq kl \ (k \in \mathbb{N}), (\lceil \frac{i}{l} \rceil - 2) \cdot l < j \leq i < N \\ 0, & \text{otherwise}, \end{cases}$$

We compress sentence information into corresponding EOUs. For dialogue generation, we retain only the first token, all EOUs, and the two most recent sentences. This task is used to train LLMs through supervised learning to adapt to replacing full dialogues with EOUs, ensuring coherent and consistent dialogue generation.

Compare

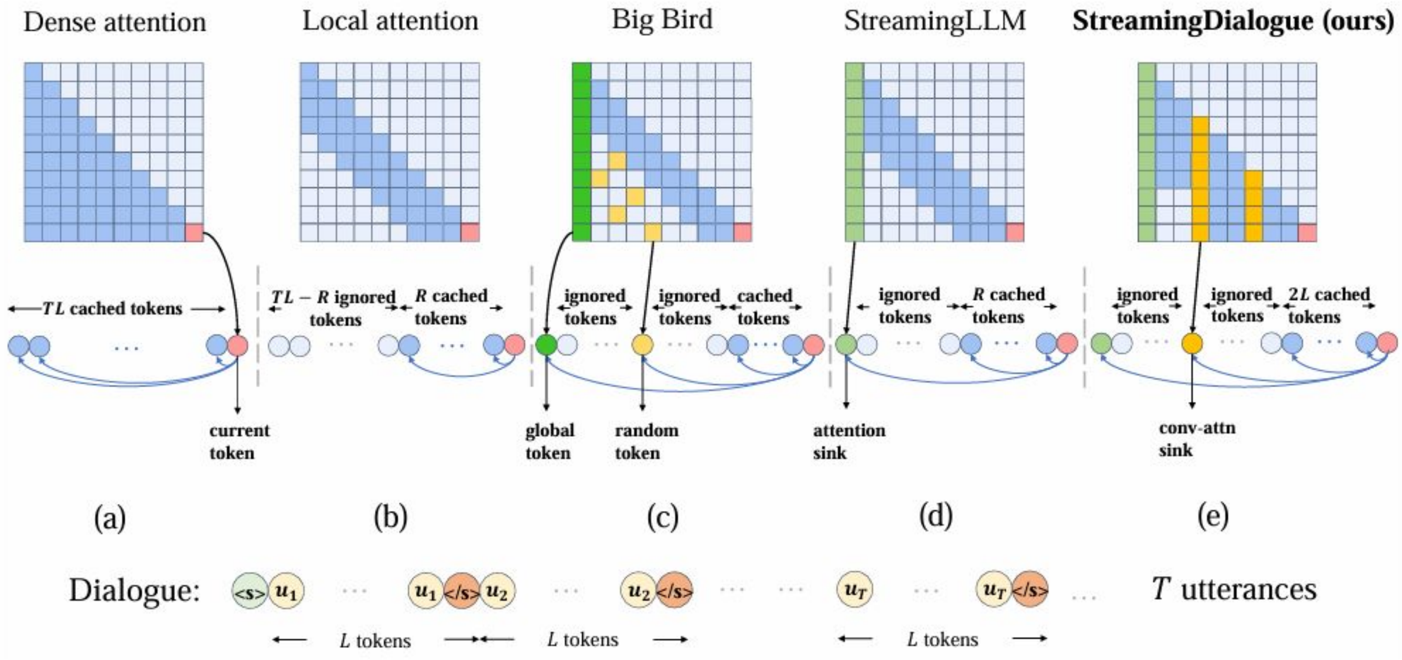


Figure 10: Attention maps' visualization of StreamingDialogue and various other methods. In a dialogue with T utterances, each averaging L tokens, dense attention caches TL tokens, local attention caches R tokens (where R is the window size), Big Bird caches global size + random size + R tokens, StreamingLLM caches $R + 1$ tokens, and StreamingDialogue requires caching up to $1 + T + 2L$ tokens.

Compared to other attention-based methods, our method not only retains historical information but also stores only a few tokens to increase speed and save resources.

Experimental Results

Table 1: Main results on the PersonaChat and MSC datasets. ↓ indicates lower values are better, while ↑ indicates the opposite. The best result for each metric is presented in bold, while the second-best one is underlined. * indicates significance ($p < 0.05$) via pairwise t-test compared to other methods. "PC" denotes PersonaChat and "StrLLM" represents StreamingLLM.

Data	Method	PPL	BLEU (%)			ROUGE (%)			Distinct (%)			USL-H (%)	Dial-M
		↓	B-avg ↑	B-1 ↑	B-2 ↑	R-1 ↑	R-2 ↑	R-L ↑	D-1 ↑	D-2 ↑	D-3 ↑	↑	↓
PC	Dense	**8.41***	13.15	49.30*	20.05*	**13.98**	**3.07**	**13.44**	**16.37***	**41.61***	**63.36***	14.21*	2.38*
PC	Local	11.59*	13.01	50.78	20.13*	13.83	2.69	13.29	12.49*	32.17*	51.12*	17.35*	**2.07**
PC	Big Bird	9.00*	12.93*	50.00*	20.52*	13.78	2.64	13.33	11.83*	32.46*	52.17*	16.95*	2.37*
PC	StrLLM	8.96	13.16	50.15	20.68	13.94	2.73	13.36	12.00*	32.64*	52.36*	17.63*	2.30*
PC	MemBART	13.15*	11.18*	46.63*	17.65*	13.11	2.56	12.78	12.86*	30.87*	48.86*	12.23*	2.49*
PC	Ours	8.71	**13.63**	**51.27**	**20.77**	13.96	3.05	13.43	14.43	37.23	58.07	**17.96**	2.10
MSC	Dense	**7.58**	**19.47**	**52.22**	**28.41**	16.93	**2.92**	15.48	**12.85***	**37.75***	**57.51***	90.11*	1.94*
MSC	Local	8.92*	13.34*	41.14*	20.44*	13.48*	1.88*	12.61*	7.89*	22.71*	35.89*	76.68*	2.15*
MSC	Big Bird	8.42*	16.54*	46.63*	24.77*	15.32*	2.34	14.15*	8.72*	25.81*	40.34*	85.30*	1.72
MSC	StrLLM	8.38*	16.76*	47.54*	25.08*	15.25*	2.44*	14.21*	9.18*	26.93*	41.62*	86.91*	**1.71**
MSC	MemBART	13.73*	17.11*	49.78*	25.82*	14.93*	2.61	13.76*	10.86*	30.55*	47.37*	85.13*	1.97*
MSC	Ours	7.99	19.33	51.49	28.12	**17.18**	2.77	**15.86**	11.54	32.58	50.27	**90.48**	1.76

Table 5: Details of dialogue datasets. We present the number of utterances (Utts.) and the average length per utterance (Avg. L) for each session in the training and test sets.

Data	Data Type	Train		Test	
		Utts.	Avg. L	Utts.	Avg. L
PersonaChat	Total	122499	13.59	14602	13.85
MSC	Session 1	59894	14.16	6572	15.47
MSC	Session 2	46420	31.44	5939	30.86
MSC	Session 3	47259	32.90	5924	32.94
MSC	Session 4	11870	32.25	5940	34.67
MSC	Session 5	-	-	5945	36.43
MSC	Total	165443	25.66	30320	29.77
Topical-Chat	Total	188378	26.76	11760	26.98
MultiWOZ	Total	113552	18.92	14744	19.23

Experimental Results

Method	B-avg ↑	R-1 ↑	R-2 ↑	D-1 ↑	D-2 ↑	USL-H ↑	Dial-M ↓
StreamingLLM	16.76	15.25	2.44	9.18	26.93	86.91	**1.71**
HRED	15.72	14.75	1.85	7.37	20.91	58.70	2.13
VHRED	17.02	15.16	1.48	5.28	14.72	59.31	2.35
Ours	**19.33**	**17.18**	**2.77**	**11.54**	**32.58**	**90.48**	1.76

Table 7: Results of the C score on the PersonaChat dataset. ↑ indicates higher values are better.

Method	Dense	Local	Big Bird	StreamingLLM	MemBART	Ours
C (%) ↑	3.10	-3.40	-4.00	-4.70	0.77	2.70

Data	Method	PPL ↓	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	Dial-M ↓
Topical-Chat	Dense	**9.49**	**15.70**	3.65	**14.88**	3.09
	Local	27.55	12.60	2.09	10.37	7.02
	Big Bird	10.36	14.21	3.55	11.79	3.01
	StreamingLLM	10.34	14.25	3.55	11.84	3.05
	MemBART	12.54	13.86	2.98	13.18	2.83
	Ours	9.80	15.46	**3.99**	14.37	**2.66**
MultiWOZ	Dense	4.51	24.79	13.93	24.67	2.27
	Local	5.38	24.26	13.47	24.15	2.45
	Big Bird	4.79	24.38	13.26	24.30	2.51
	StreamingLLM	4.76	23.66	13.09	23.41	2.47
	MemBART	5.36	20.05	12.41	19.94	2.37
	Ours	**4.34**	**25.26**	**14.27**	**25.20**	**2.25**

Experimental Results

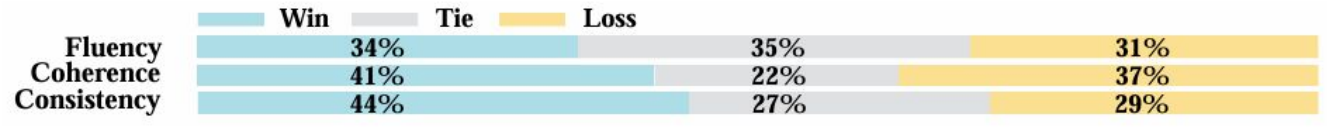


Figure 3: Fluency, coherence, and consistency in human evaluations: ours vs StreamingLLM.

Table 2: Ablation results on MSC with different learning strategies. "Base" denotes the model fine-tuned without SMR and LMR learning.

Model	PPL	BLEU-avg	ROUGE-L	Distinct-3
Ours	7.99	19.33	15.86	50.27
Base	8.21	17.32	10.25	46.15
LMR	8.01	18.87	15.66	49.44
SMR	8.40	18.25	15.24	48.57

Table 3: Results under the non-training setting on the MSC test set.

Model	Method	BLEU-avg	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	ROUGE-L
Llama-2-7B-Chat	StreamingLLM	20.16	51.18	29.99	15.90	1.92	14.26
	Ours	20.19	51.55	30.03	16.46	2.11	15.00
Llama-3-8B-Instruct	StreamingLLM	16.48	39.68	24.63	16.88	1.93	15.47
	Ours	16.77	40.10	24.88	17.11	2.01	15.85
Mistral-7B	StreamingLLM	12.75	42.86	19.99	12.58	1.83	11.73
	Ours	13.33	44.08	20.65	13.40	1.98	12.58

Experimental Results

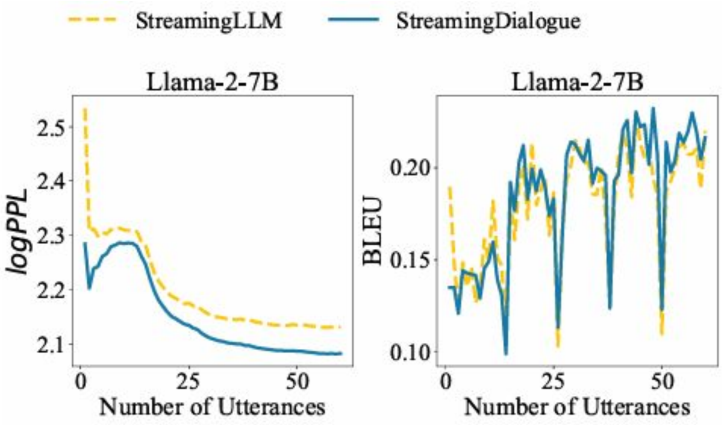


Figure 4: Average perplexity and BLEU for StreamingLLM and StreamingDialogue on the MSC test set across varying utterance counts.

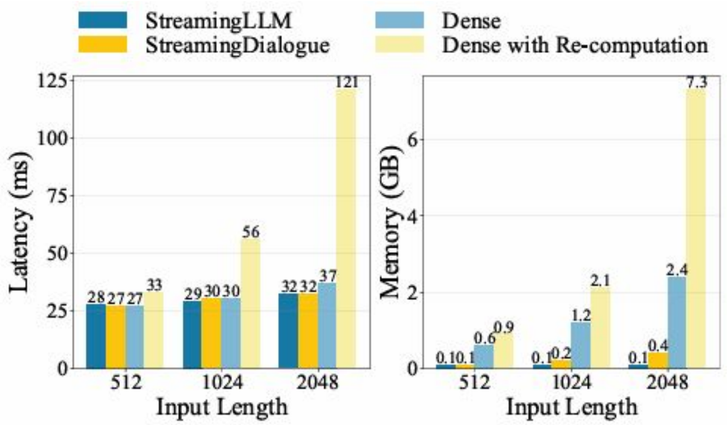


Figure 5: Per-token latency and memory usage by method on MSC for varying input lengths, with memory reported as total minus fixed.

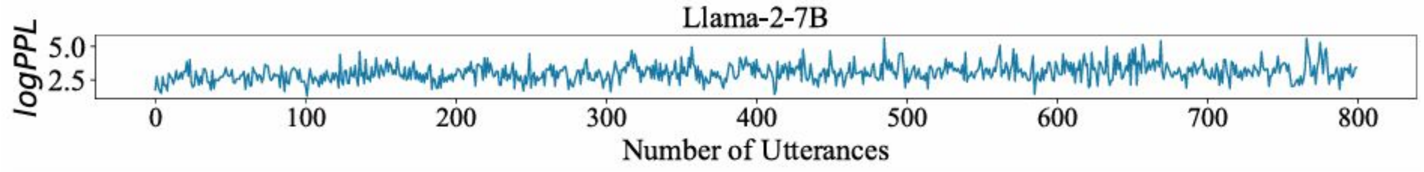


Figure 6: The perplexity for StreamingDialogue under the concatenated MSC test set, evaluating approximately 25K tokens.

Experimental Results

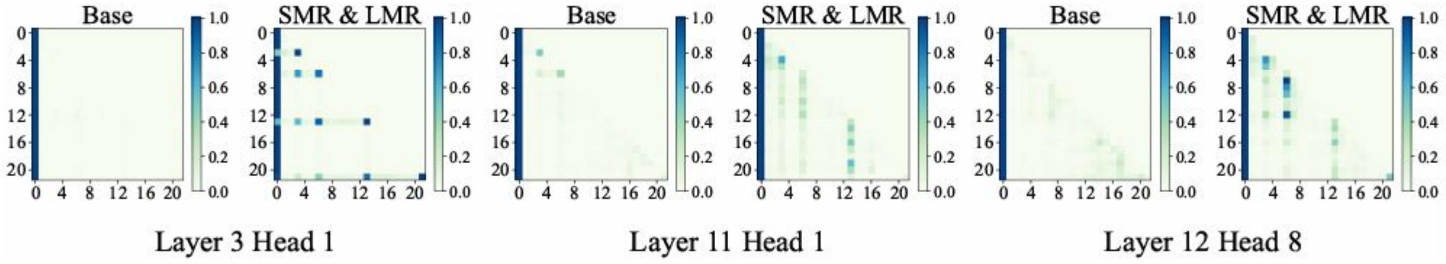


Figure 7: Comparison of attention maps before and after learning. “Base” denotes Llama-2-7B, while “SMR & LMR” represents the model obtained post co-training with SMR and LMR on Llama-2-7B. The “</s>” positions in the encoded sentences are: 3, 6, 13, and 21.

Table 4: Dialogue reconstruction performance.

BLEU-avg	BLEU-1	BLEU-2	ROUGE-1	ROUGE-L
68.02	89.19	76.83	76.79	72.94

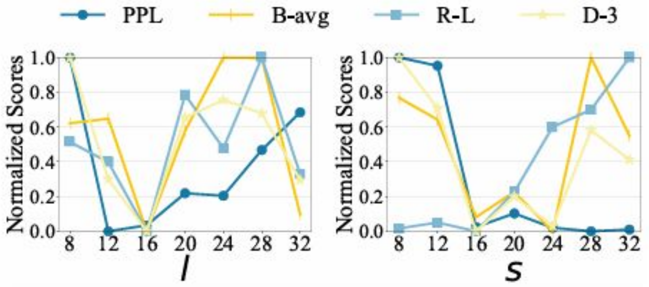


Figure 9: Normalized performance scores (PPL, B-avg, R-L, and D-3) on MSC for various l with s fixed at 28 and various s with l fixed at 24.

Experimental Results

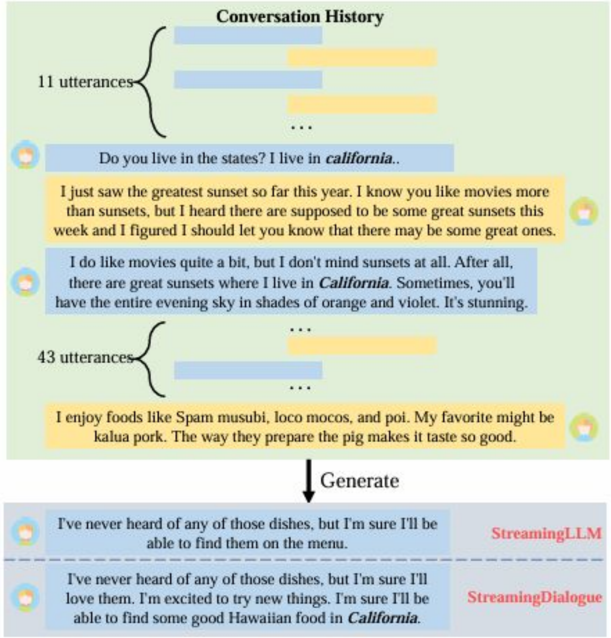


Figure 8: The generated dialogues by StreamingLLM and StreamingDialogue for the same input dialogue history from an MSC episode, with an average utterance length of $L = 32$ tokens. Bold italic indicates key information in the dialogue.

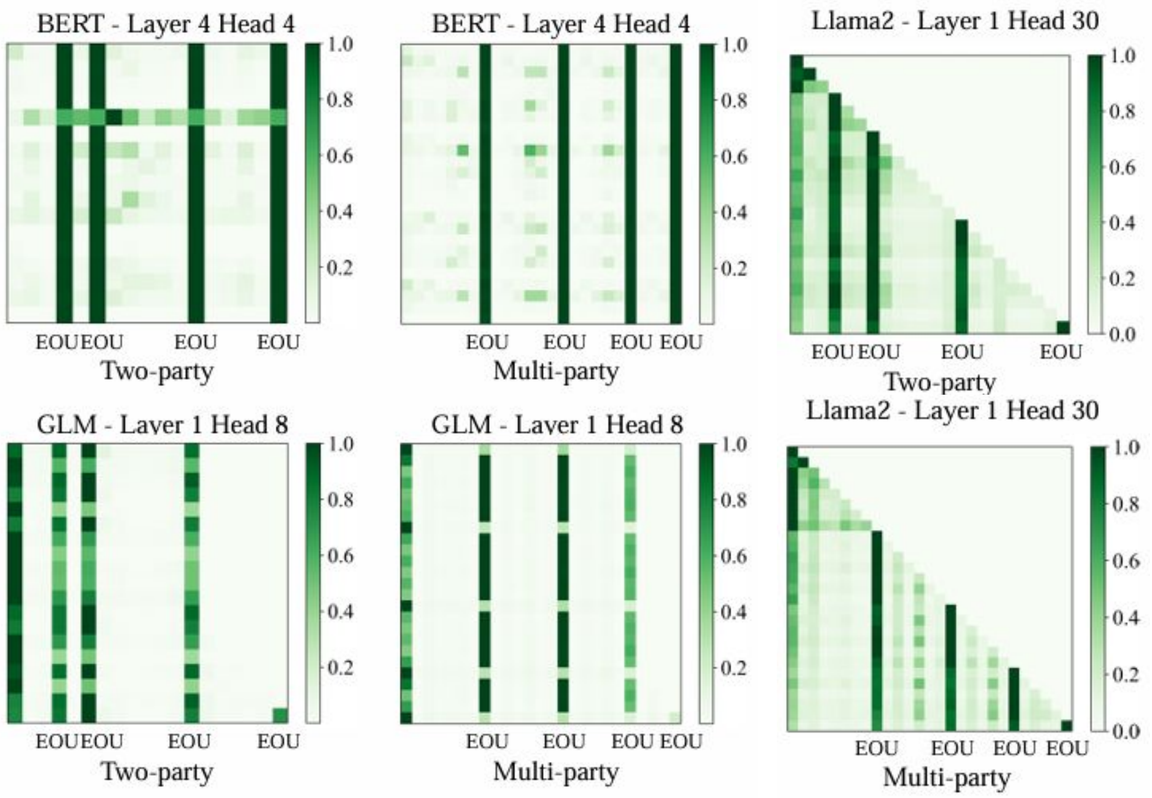


Figure 11: Attention maps under different settings

Experimental Results

Table 11: The results of integrating the belief states on the MultiWOZ dataset.

Method	PPL	BLEU-avg	BLEU-1	BLEU-2	Distinct-1	Distinct-2	Distinct-3
Dense	1.92	25.56	48.33	29.14	3.74	6.86	8.89
StreamingLLM	2.19	25.70	47.53	29.21	4.48	9.09	12.60
Ours	1.98	25.77	48.58	29.38	5.30	10.03	13.60

Table 12: Results on the Topical-Chat and Persona-Chat datasets under the setting of treating each sentence of the grounding knowledge/persona profiles as an utterance.

Data	Method	PPL	Distinct-2	Distinct-3	Dial-M
PersonaChat	Dense	7.19	43.56	66.27	2.53
	StreamingLLM	8.36	33.17	53.58	2.47
	Ours	7.60	39.16	61.06	2.36
Topical-Chat	Dense	3.24	39.07	57.64	4.32
	StreamingLLM	8.31	16.87	23.56	3.72
	Ours	3.20	31.47	49.10	2.57

Table 13: Results on the Topical-Chat and Persona-Chat datasets under the setting of treating the grounding knowledge/persona profiles as a prompt.

Data	Method	PPL	Distinct-2	Distinct-3	Dial-M
PersonaChat	Dense	7.93	44.26	66.63	2.48
	StreamingLLM	7.99	36.40	57.44	2.91
	Ours	7.67	37.82	58.93	2.57
Topical-Chat	Dense	11.64	36.98	54.96	4.60
	StreamingLLM	30.37	26.07	34.26	3.61
	Ours	10.21	32.16	50.41	2.97

Experimental Results

Analysis of EoU tokens' information aggregation capability

Examples of prompt formats are as follows, where the "keywords" will be replaced with specific content.

1. "template": "A and B went to PLACE today.</s>They had a great time.</s>Who did A go to PLACE with today?</s>",
   "keywords": "A": "person", "B": "person", "PLACE": "place",
   "answer key": "B"

2. "template": "B made A's favorite food, FOOD, today.</s>A was delighted.</s>What food did B make for A today?</s>",
   "keywords": "A": "person", "B": "person", "FOOD": "food",
   "answer key": "FOOD"

3. "template": "A was doing ACTIVITY when B called.</s>A had to stop and answer the call.</s>What was A doing when B called?</s>",
   "keywords": "A": "person", "B": "person", "ACTIVITY": "activity",
   "answer key": "ACTIVITY"

4. "template": "A bought a new ITEM today.</s>B was impressed by A's purchase.</s>What item did A buy today?</s>",
   "keywords": "A": "person", "B": "person", "ITEM": "item",
   "answer key": "ITEM"

5. "template": "A participated in an EVENT today.</s>B cheered them on.</s>What event did A participate in?</s>",
   "keywords": "A": "person", "B": "person", "EVENT": "event",
   "answer key": "EVENT"

Accuracy: ***68%***

Gaoling School of Artificial Intelligence
Renmin University of China

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you!

Feel free to contact us for further discussion

Jia-Nan Li
lijianan@ruc.edu.cn

WeChat for Jia-Nan Li

Github Homepage for Jia-Nan Li