Microsoft

ICLR
International Conference On Learning Representations

SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents.

Zhuoshi Pan[1], Qianhui Wu[2], Huiqiang Jiang[2], Xufang Luo[2], Hao Cheng[2], Dongsheng Li[2], Yuqing Yang[2], Chin-Yew Lin[2], H. Vicky Zhao[1], Lili Qiu[2], Jianfeng Gao[2]

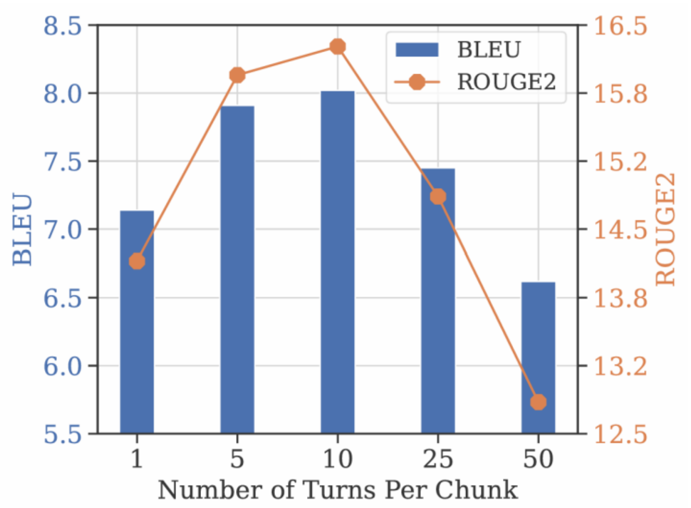
1: Tsinghua University
2: Microsoft Corporation

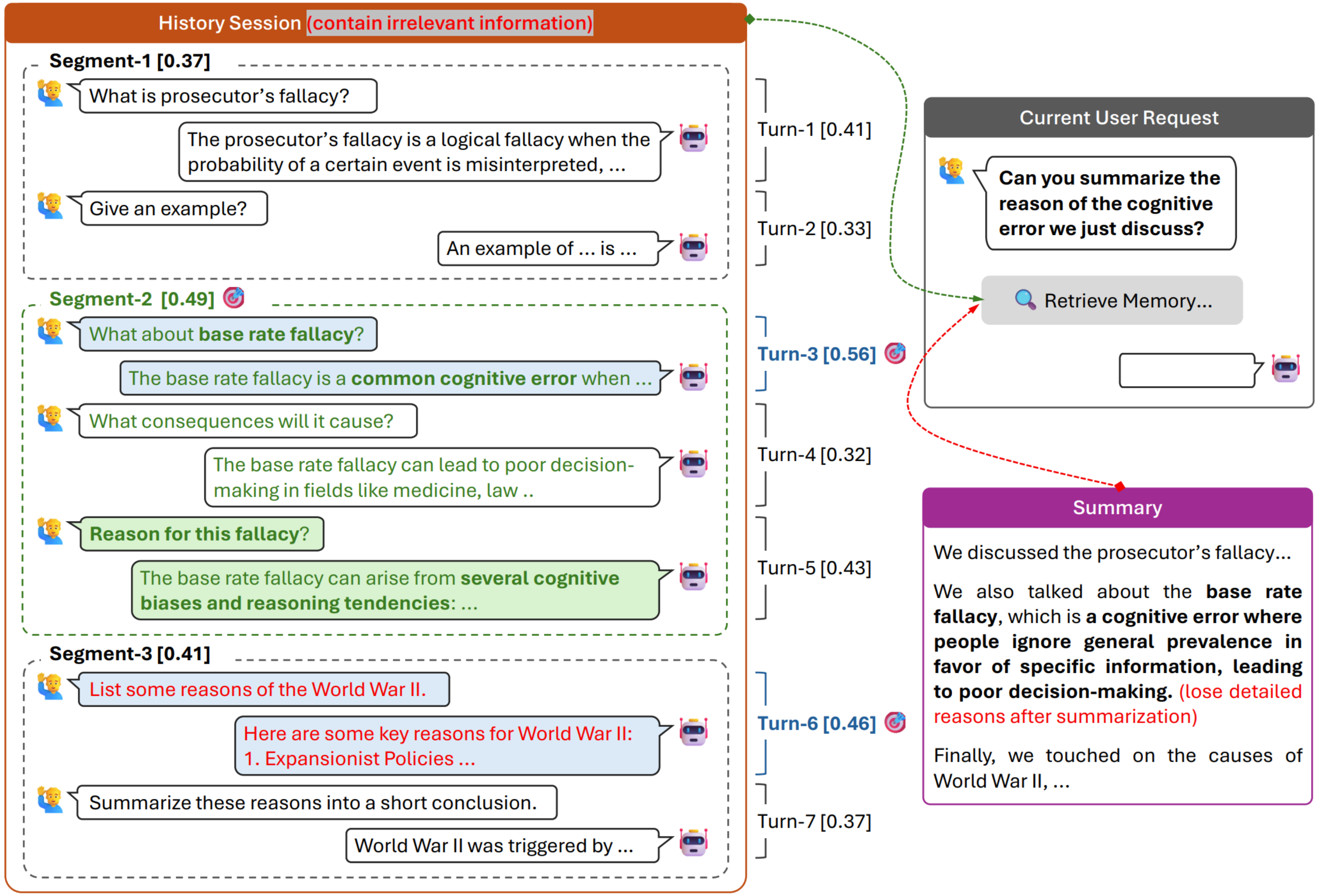
https://aka.ms/secom

Microsoft Research

Challenges Encountered in RAG for Conversational Agents

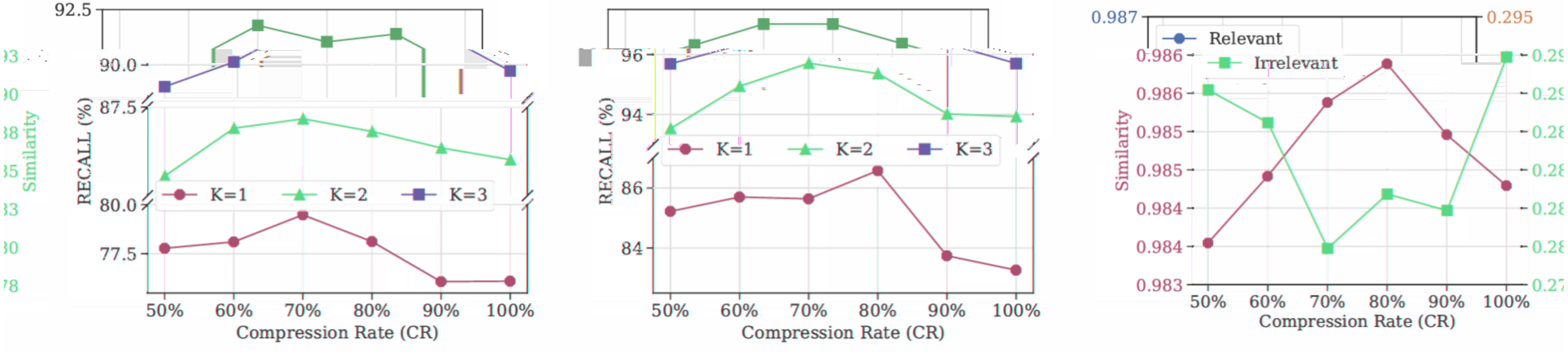
- **Challenges 1: the granularity of memory unit matters**
  - Turn-level: too fine-grained ✖
  - Session-level: too coarse-grained ✖
  - Summarization-based methods: suffer from information loss ✖





Challenges Encountered in RAG for Conversational Agents

- **Challenges 2: the redundancy in natural language impairs the retrieval system**
  - Decreasing the retrieval recall.
  - Complicating the extraction of key information



(a) Retrieval recall v.s. compression rate: $\frac{\#\text{tokens after compression}}{\#\text{tokens before compression}}$. K: number of retrieved segments. Retriever: BM25.

(b) Retrieval recall v.s. compression rate: $\frac{\#\text{tokens after compression}}{\#\text{tokens before compression}}$. K: number of retrieved segments. Retriever: MPNet.

(c) Similarity between the query and different dialogue segments. Blue: releva segments. Orange: irrelevant segments. Retriever: MPNet.

SeCom

A system constructs memory bank at segment-level and applies compression-based denoising on memory unit

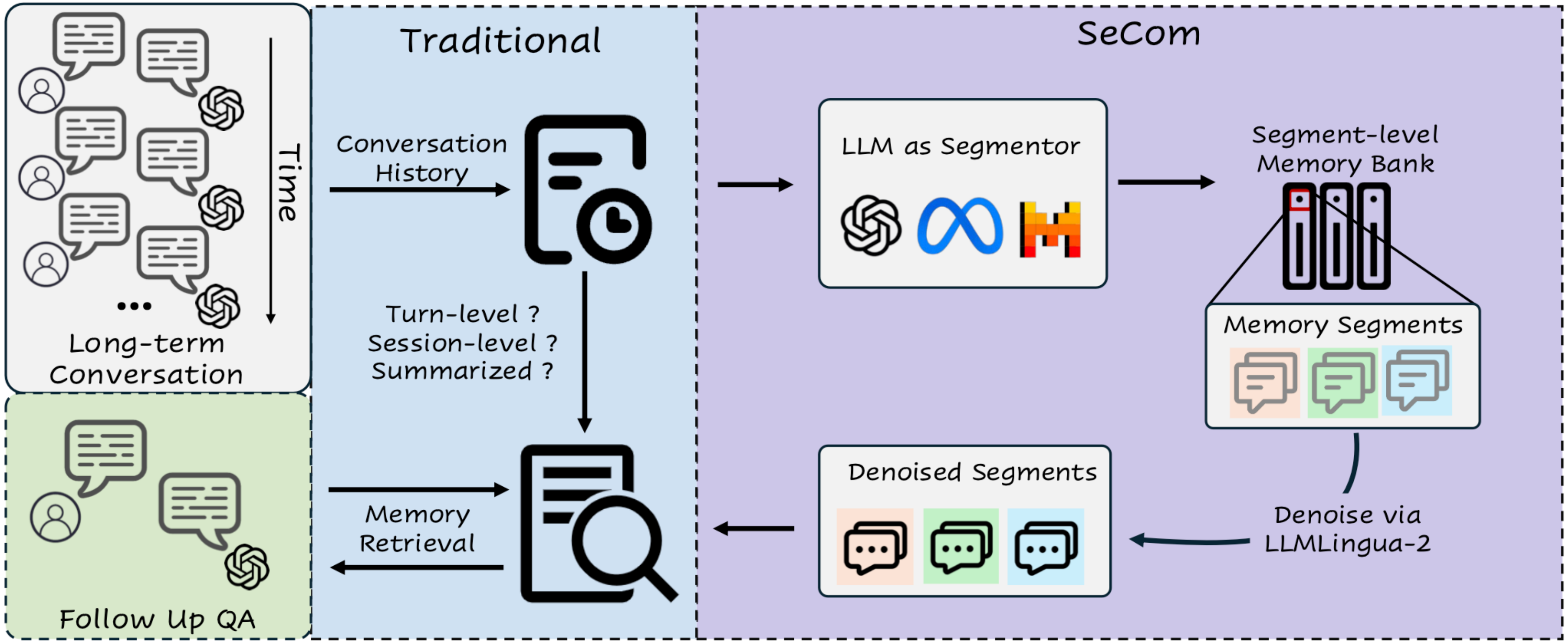
Conversation Segmentation Model:

✓ Segments a conversation session into several segments.

✓ Lightweight models, such as ***Mistral-7B*** and even ***RoBERTa-scale models*** can perform segmentation well.

Compression-Based Memory Denoising:

✓ Employ ***LLMLingua-2*** to compress the memory unit before retrieval.



Experiments | Main Results

SeCom outperforms all baseline approaches and exhibits greater robustness of the retrieval system.

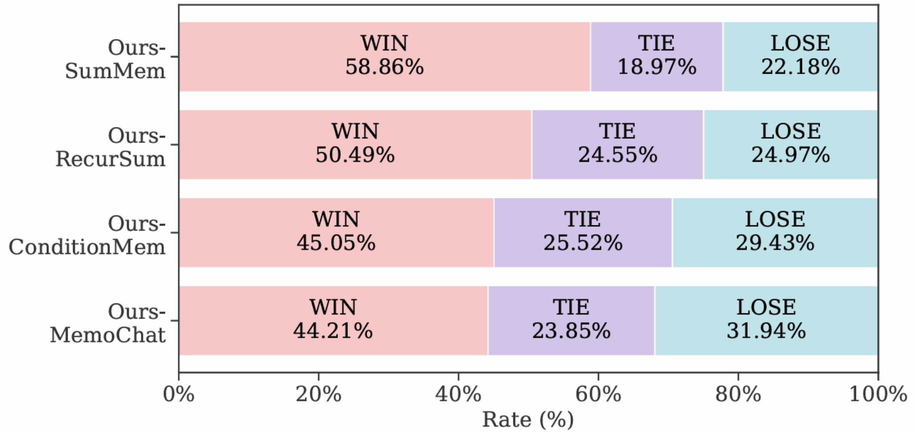
Experimental Setup

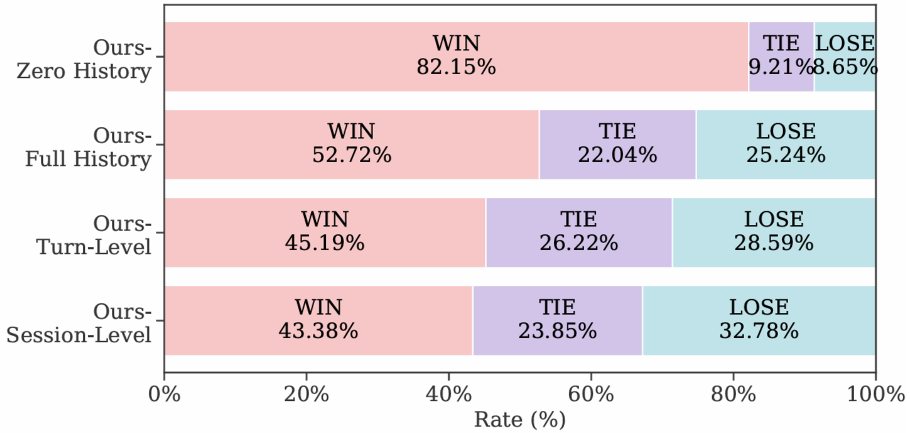
- **Datasets**: LOCOMO, Long-MT-Bench+.
- **Segmentation Models**: GPT-4, Mistral-7B-Instruct-v0.3, RoBERTa-based model
- **Response Models**: GPT-35-Turbo, Mistral-7B-Instruct-v0.3
- **Retrievers**: BM25, MPNet-based.
- **Baselines**: Full History, Turn-level, Session-level and four strong baselines.

Methods	QA Performance						Context Length	
	GPT4Score	BLEU	Rouge1	Rouge2	RougeL	BERTScore	# Turns	# Tokens
LOCOMO								
Zero History	24.86	1.94	17.36	3.72	13.24	85.83	0.00	0
Full History	54.15	6.26	27.20	12.07	22.39	88.06	210.34	13,330
Turn-Level (BM25)	65.58	7.05	29.12	13.87	24.21	88.44	49.82	3,657
Turn-Level (MPNet)	57.99	6.07	26.61	11.38	21.60	88.01	54.77	3,288
Session-Level (BM25)	63.16	7.45	29.29	14.24	24.29	88.33	55.88	3,619
Session-Level (MPNet)	51.18	5.22	24.23	9.33	19.51	87.45	53.88	3,471
SumMem	53.87	2.87	20.71	6.66	16.25	86.88	-	4,108
RecurSum	56.25	2.22	20.04	8.36	16.25	86.47	-	400
ConditionMem	65.92	3.41	22.28	7.86	17.54	87.23	-	3,563
MemoChat	65.10	6.76	28.54	12.93	23.65	88.13	-	1,159
SeCom (BM25, GPT4-Seg)	**71.57**	**8.07**	**31.40**	**16.30**	**26.55**	**88.88**	55.52	3,731
SeCom (MPNet, GPT4-Seg)	69.33	7.19	29.58	13.74	24.38	88.60	55.51	3,716
SeCom (MPNet, Mistral-7B-Seg)	66.37	6.95	28.86	13.21	23.96	88.27	55.80	3,720
SeCom (MPNet, RoBERTa-Seg)	61.84	6.41	27.51	12.27	23.06	88.08	56.32	3,767
Long-MT-Bench+								
Zero History	49.73	4.38	18.69	6.98	13.94	84.22	0.00	0
Full History	63.85	7.51	26.54	12.87	20.76	85.90	65.45	19,287
Turn-Level (BM25)	82.85	11.52	32.84	17.86	26.03	87.03	3.00	1,047
Turn-Level (MPNet)	84.91	12.09	34.31	19.08	**27.82**	86.49	3.00	909
Session-Level (BM25)	81.27	11.85	32.87	17.83	26.82	87.32	13.35	4,118
Session-Level (MPNet)	73.38	8.89	29.34	14.30	22.79	86.61	13.43	3,680
SumMem	63.42	7.84	25.48	10.61	18.66	85.70	-	1,651
RecurSum	62.96	7.17	22.53	9.42	16.97	84.90	-	567
ConditionMem	63.55	7.82	26.18	11.40	19.56	86.10	-	1,085
MemoChat	85.14	12.66	33.84	19.01	26.87	87.21	-	1,615
SeCom (BM25, GPT4-Seg)	86.67	12.74	33.82	18.72	26.87	87.37	2.87	906
SeCom (MPNet, GPT4-Seg)	**88.81**	**13.80**	**34.63**	**19.21**	27.64	**87.72**	2.77	820
SeCom (MPNet, Mistral-7B-Seg)	86.32	12.41	34.37	19.01	26.94	87.43	2.85	834
SeCom (MPNet, RoBERTa-Seg)	81.52	11.27	32.66	16.23	25.51	86.63	2.96	841

Pairwise Comparison & Human Evaluation

☐ Pairwise Comparison (GPT-4 Judge)





☐ Human Evaluation

Methods	Coherence	Consistency	Memorability	Engagingness	Humanness	Average
Full-History	1.55	1.11	0.43	0.33	1.85	1.05
Sentence-Level	1.89	1.20	1.06	0.78	2.00	1.39
Session-Level	1.75	1.25	0.98	0.80	1.92	1.34
ConditionMem	1.58	1.08	0.57	0.49	1.77	1.10
MemoChat	2.05	1.25	1.12	0.86	**2.10**	1.48
COMEDY	**2.20**	1.28	1.20	0.90	1.97	1.51
SECOM (Ours)	2.13	**1.34**	**1.28**	**0.94**	2.06	1.55

Segmentation Evaluation

LLM-based segmentation outperforms unsupervised baselines.

Experimental Setup

- **Datasets**: DialSeg711, TIAGE and SuperDialSeg.

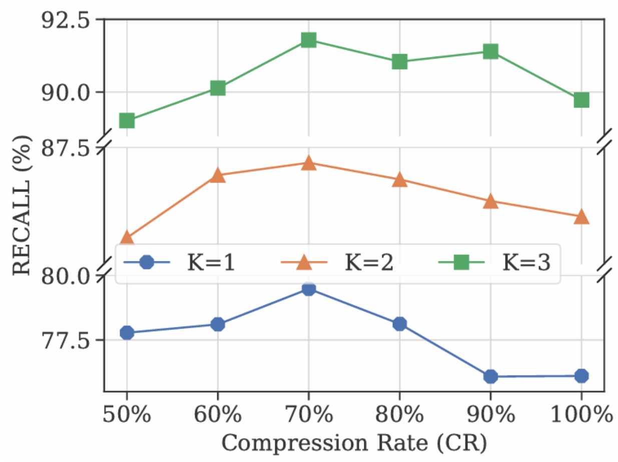
				Dialseg711				SuperDialSeg				TIAGE
WD↓	F1↑	Score↑		Pk↓	WD↓	F1↑	Score↑	Pk↓	WD↓	F1↑	Score↑	Pk↓
			Unsupervised									
0.571	0.366	0.419	BayesSeg	0.306	0.350	0.556	0.614	0.433	0.593	0.438	0.463	0.486
0.488	0.204	0.363	TextTiling	0.470	0.493	0.245	0.382	0.441	0.453	0.388	0.471	0.469
0.515	0.238	0.366	GraphSeg	0.412	0.442	0.392	0.483	0.450	0.454	0.249	0.398	0.496
0.511	0.236	0.369	TextTiling+Glove	0.399	0.438	0.436	0.509	0.519	0.524	0.353	0.416	0.486
0.556	0.218	0.340	TextTiling+[CLS]	0.419	0.473	0.351	0.453	0.493	0.523	0.277	0.385	0.521
0.439	0.285	0.426	TextTiling+NSP	0.347	0.360	0.347	0.497	0.512	0.521	0.208	0.346	0.425
0.506	0.181	0.341	GreedySeg	0.381	0.410	0.445	0.525	0.490	0.494	0.365	0.437	0.490
0.420	0.427	0.509	CSM	0.278	0.302	0.610	0.660	0.462	0.467	0.381	0.458	0.400
-	-	-	DialSTART †	0.178	0.198	-	-	-	-	-	-	-
0.401	**0.596**	**0.607**	**Ours**	**0.093**	**0.103**	**0.888**	**0.895**	**0.277**	**0.289**	**0.758**	**0.738**	**0.363**

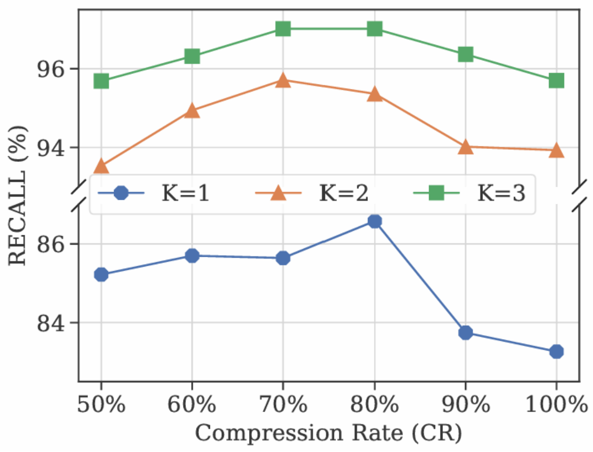
Ablation Study

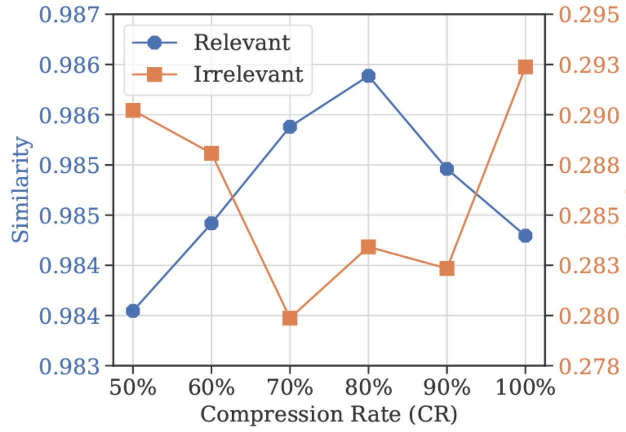
Removing the Compression-based denoising degrades the performance.

Methods	LOCOMO				Long-MT-Bench+			
	GPT4Score	BLEU	Rouge2	BERTScore	GPT4Score	BLEU	Rouge2	BERTScore
SeCom	**69.33**	**7.19**	**13.74**	**88.60**	**88.81**	**13.80**	**19.21**	**87.72**
– Denoise	59.87	6.49	12.11	88.16	87.51	12.94	18.73	87.44

Reason: (a) improving the retrieval recall (b) increasing the similarity between the query and relevant segments while decreasing the similarity with irrelevant ones.







You can find more details in

aka.ms/SeCom