SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents

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

Microsoft Corporation

Motivation & Methodology

- **Background:** Long-term, open-domain conversations over multiple sessions conversations challenges the LLM-powered conversational agent[1,2], as they require the system to retain past events and user preferences to deliver coherent and personalized responses.
- **Findings:**
  - We first systematically investigate the impact of memory granularities on *retrieval augmented conversational agents*, and find that commonly used turn-level[3], session-level[4], and summarization-based methods[2,5] all exhibit limitations.

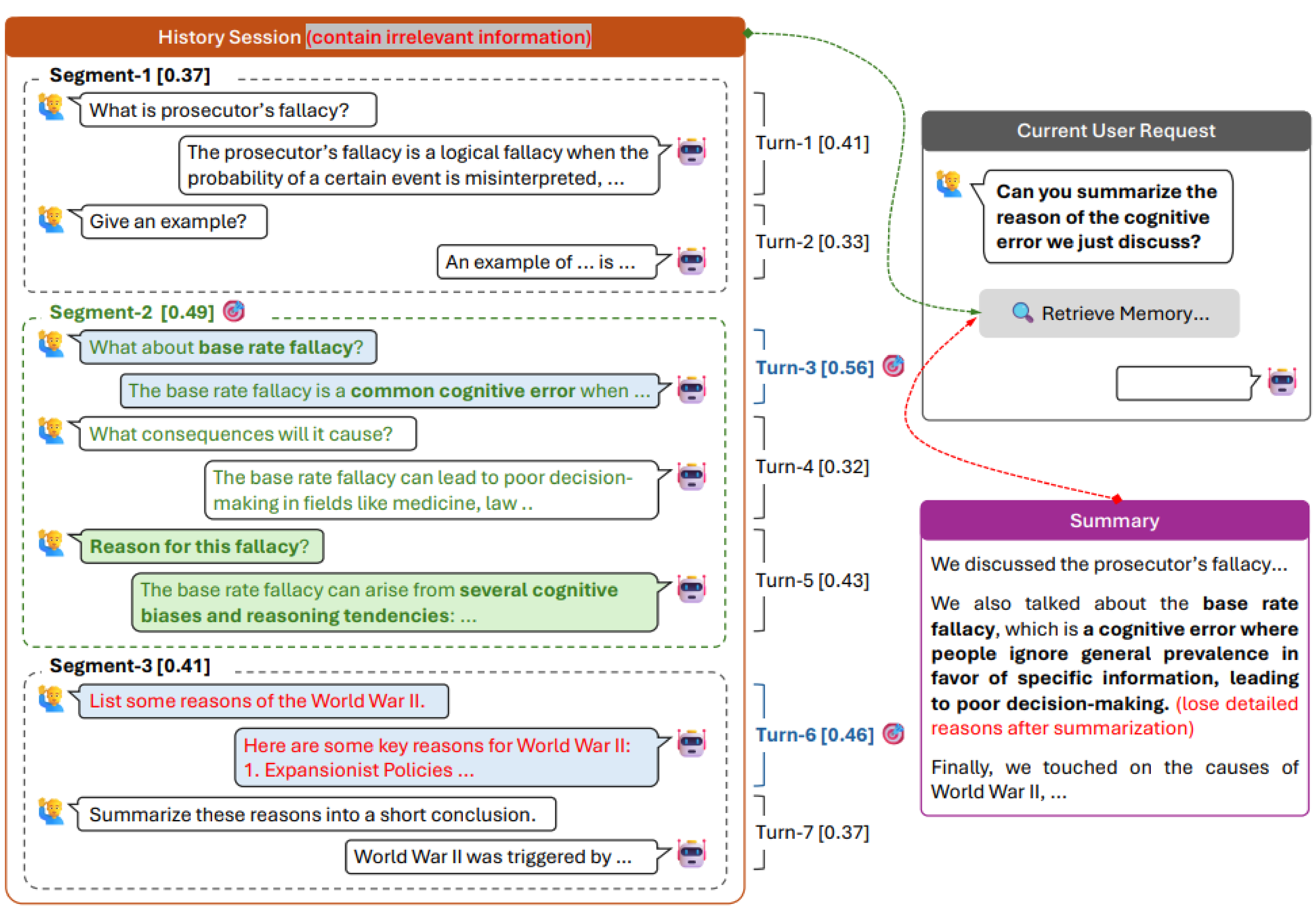


Figure: Illustration of retrieval augmented conversation agent with different memory granularities. Ours (segment-level memory) can better capture topically coherent unit.

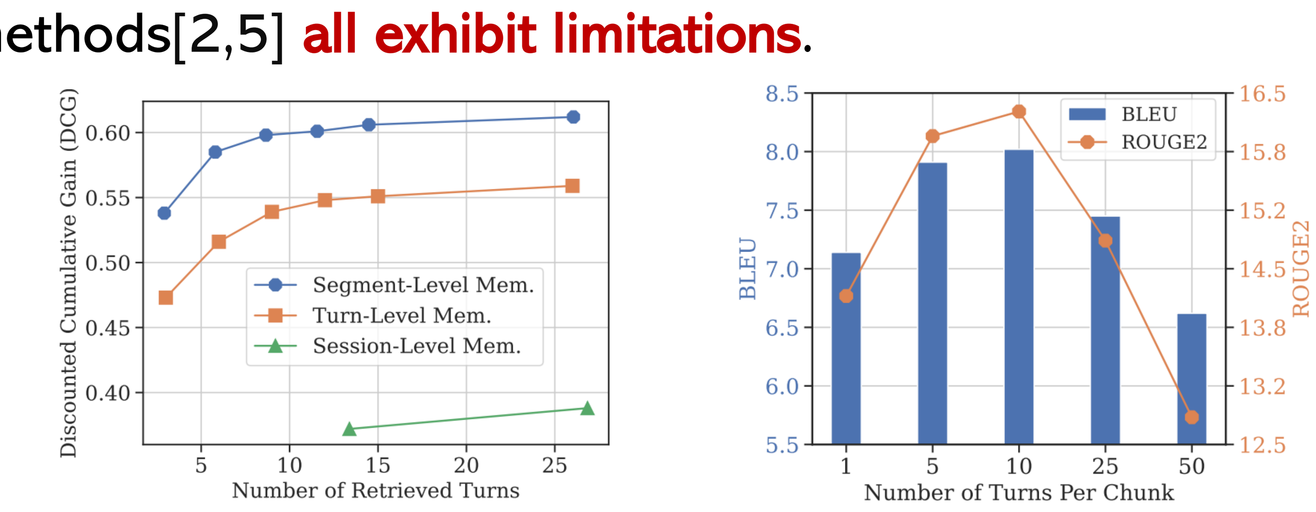


Figure: memory granularity impacts (1) retrieval accuracy and (2) response quality.

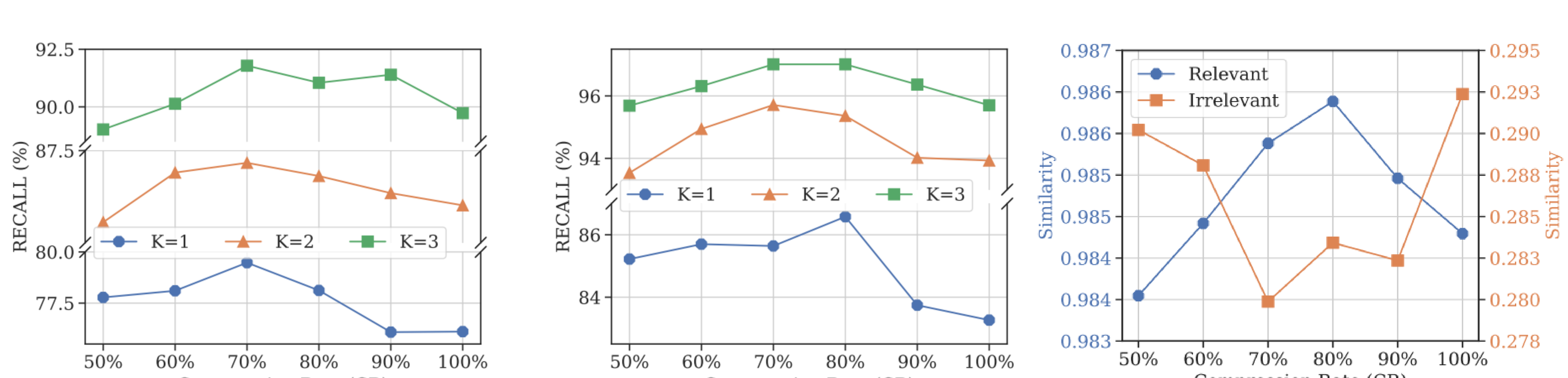
Turn-level memory is **too fine-grained**, leading to **fragmentary and incomplete context** and misses essential interaction turns.

Session-level memory is **too coarse-grained**, **containing too much irrelevant information**, which distracts both the retrieval module and the LLM.

Summary-based methods **suffer from information loss** that occurs during summarization.

Our SeCom can better capture topically coherent units, balancing 1) **including relevant information** and 2) **excluding irrelevant content**.

  - **Redundancy** in long-term conversation **acts as noise, hindering accurate memory retrieval**[6].



Redundancy in long-term conversation **acts as noise for retrieval systems** and prompt compression helps.

Prompt compression method like LLMLingua-2 removes such redundancy, **increasing the similarity** between the **query and relevant segments** and **decreasing the similarity with irrelevant ones**.

- **Methodology:**
  - Introducing a conversation segmentation model that **partitions long-term conversations into topically coherent segments**, constructing the memory bank at the **segment level**: $\mathcal{H} = \{\mathbf{c}_i\}_{i=1}^{C} \rightarrow \mathcal{M} = \{m_i\}_{i=1}^{M}$
  - **Removing redundancy from memory units** prior to retrieval by leveraging **prompt compression** method[7]: $\{m_n \in \mathcal{M}\}_{n=1}^{N} \leftarrow f_R(u^*, f_{Comp}(\mathcal{M}), N)$.
  - Integrating these two technologies into a unified system, **SeCom**, towards better personalized conversational agents.

Experiments & Discussion

- **Overall Effectiveness:** SECOM, which constructs memory bank at segment level, outperforms SOTA baseline approaches. Moreover, more light-weight segmentation models remain effective.

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 (MPNet)	57.99	6.07	26.61	11.38	21.60	88.01	54.77	3,288
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 (RoBERTa-Seg)	61.84	6.41	27.51	12.27	23.06	88.08	56.32	3,767
SECOM (Mistral-7B-Seg)	66.37	6.95	28.86	13.21	23.96	88.27	55.80	3,720
SECOM (GPT-4-Seg)	**69.33**	**7.19**	**29.58**	**13.74**	**24.38**	**88.60**	55.51	3,716
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 (MPNet)	84.91	12.09	34.31	19.08	**27.82**	86.49	3.00	909
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 (RoBERTa-Seg)	81.52	11.27	32.66	16.23	25.51	86.63	2.96	841
SECOM (Mistral-7B-Seg)	86.32	12.41	34.37	19.01	26.94	87.43	2.85	834
SECOM (GPT-4-Seg)	**88.81**	**13.80**	**34.63**	**19.21**	27.64	**87.72**	2.77	820

- **Effectiveness of the Conversation Segmentation Model**: our segmentation model is well suited for unsupervised scenarios.

Methods	Dialseg711				SuperDialSeg				TIAGE			
	Pk↓	WD↓	F1↑	Score↑	Pk↓	WD↓	F1↑	Score↑	Pk↓	WD↓	F1↑	Score↑
Unsupervised Baselines												
BayesSeg	0.306	0.350	0.556	0.614	0.433	0.593	0.438	0.463	0.486	0.571	0.366	0.419
TextTiling	0.470	0.493	0.245	0.382	0.441	0.453	0.388	0.471	0.469	0.488	0.204	0.363
GraphSeg	0.412	0.442	0.392	0.483	0.450	0.454	0.249	0.398	0.496	0.515	0.238	0.366
TextTiling+Glove	0.399	0.438	0.436	0.509	0.519	0.524	0.353	0.416	0.486	0.511	0.236	0.369
TextTiling+[CLS]	0.419	0.473	0.351	0.453	0.493	0.523	0.277	0.385	0.521	0.556	0.218	0.340
TextTiling+NSP	0.347	0.360	0.347	0.497	0.512	0.521	0.208	0.346	0.425	0.439	0.285	0.426
GreedySeg	0.381	0.410	0.445	0.525	0.490	0.494	0.365	0.437	0.490	0.506	0.181	0.341
CSM	0.278	0.302	0.610	0.660	0.462	0.467	0.381	0.458	0.400	0.420	0.427	0.509
Transfer-learning Based Baselines												
Training Set	**Train on TIAGE**				**Train on TIAGE**				**Train on SuperDialSeg**			
TextSeg_{dial}	0.476	0.491	0.182	0.349	0.552	0.570	0.199	0.319	0.489	0.508	0.266	0.384
BERT	0.441	0.411	0.005	0.297	0.511	0.513	0.043	0.266	0.492	0.526	0.226	0.359
RoBERTa	0.197	0.210	0.650	0.723	0.434	0.436	0.276	0.420	0.401	0.418	0.373	0.482
LLM-based Segmentation Model (Zero-Shot)												
Ours	**0.093**	**0.103**	**0.888**	**0.895**	**0.277**	**0.289**	**0.758**	**0.738**	**0.363**	**0.401**	**0.596**	**0.607**

- **Ablation on Compression Denoising:** removing compression-based denoising mechanism leads to performance drop, particularly on the long-conversation benchmark LOCOMO.

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

[1] Maharana et al., Evaluating very long-term conversational memory of llm agents. ACL 2024.
[2] Chen et al., Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversation, 2024.
[3] Yuan et al., Evolving large language model assistant with long-term conditional memory. 2024.
[4] Wang et al., Recursively summarizing enables long-term dialogue memory in large language models. 2024.
[5] Xu et al., Beyond goldfish memory: Long-term open-domain conversation. ACL 2022.
[6] Ma et al., Simple and effective unsupervised redundancy elimination to compress dense vectors for passage retrieval. EMNLP, 2021.
[7] Pan et al., LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression. ACL 2024.