NEURAL INFORMATION PROCESSING SYSTEMS
GenAI Workshop

Demo: An Exploration of LLM-Guided Conversation in Reminiscence Therapy

Jinhao Duan[1*], Xinyu Zhao[2*], Zhuoxuan Zhang[3*], Eunhye Grace Ko[4], Lily Boddy[4], Chenan Wang[1], Tianhao Li[4], Alexander Rasgon[4], Junyuan Hong[4], Min Kyung Lee[4], Chenxi Yuan[5], Qi Long[6], Ying Ding[4], Tianlong Chen[2], Kaidi Xu[1]

[1]Drexel University [2]UNC Chapel Hill [3]Brown University [4]UT Austin
[5]New Jersey Institute of Technology [6]University of Pennsylvania

What is LLM-Guided Conversation?

NEURAL INFORMATION PROCESSING SYSTEMS
GenAI Workshop

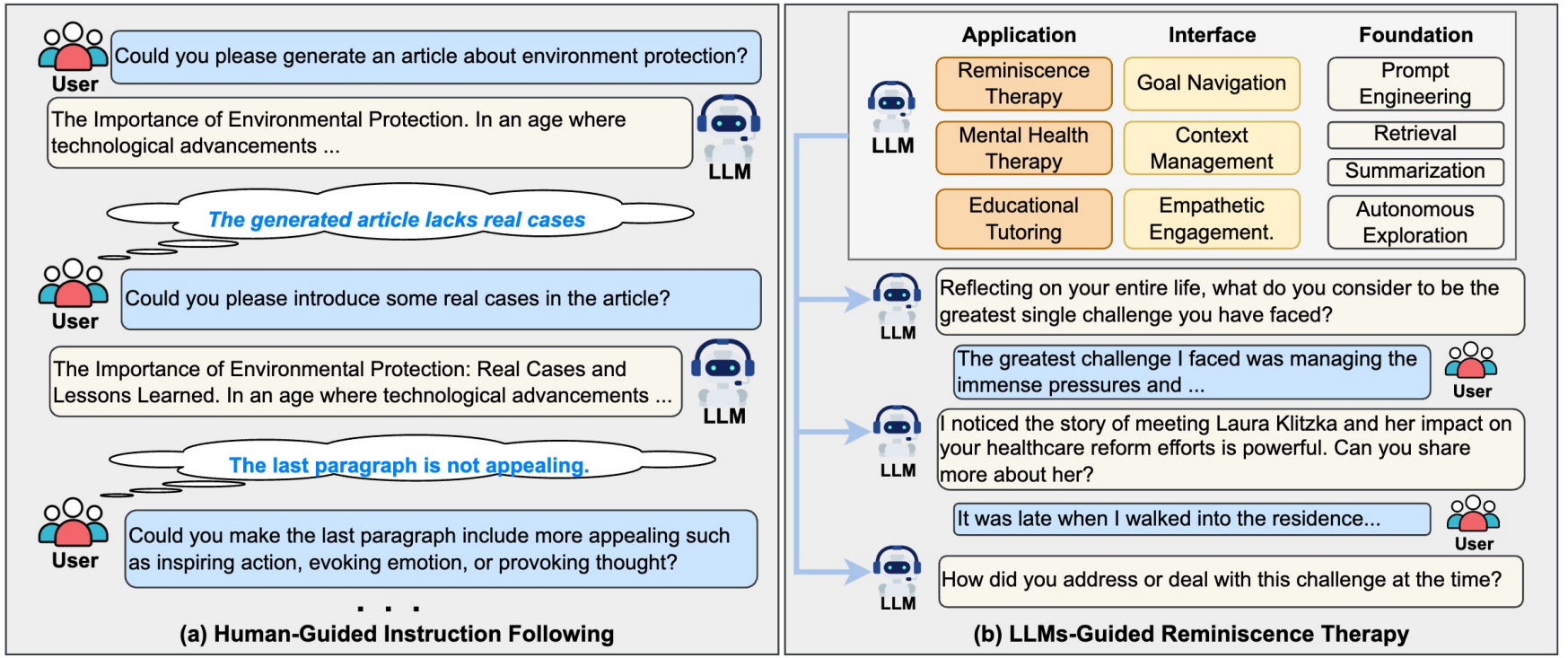


Figure 1: Comparison between human-guided instruction following and LLM-guided conversation. (a) Human-Guided: Human dominates the conversation, providing feedback and instruction to LLMs. (b) LLM-Guided: LLMs navigate the goal by automatically extrapolating reminisce questions.

GuideLLM: Functions and Architecture

NEURAL INFORMATION PROCESSING SYSTEMS
GenAI Workshop

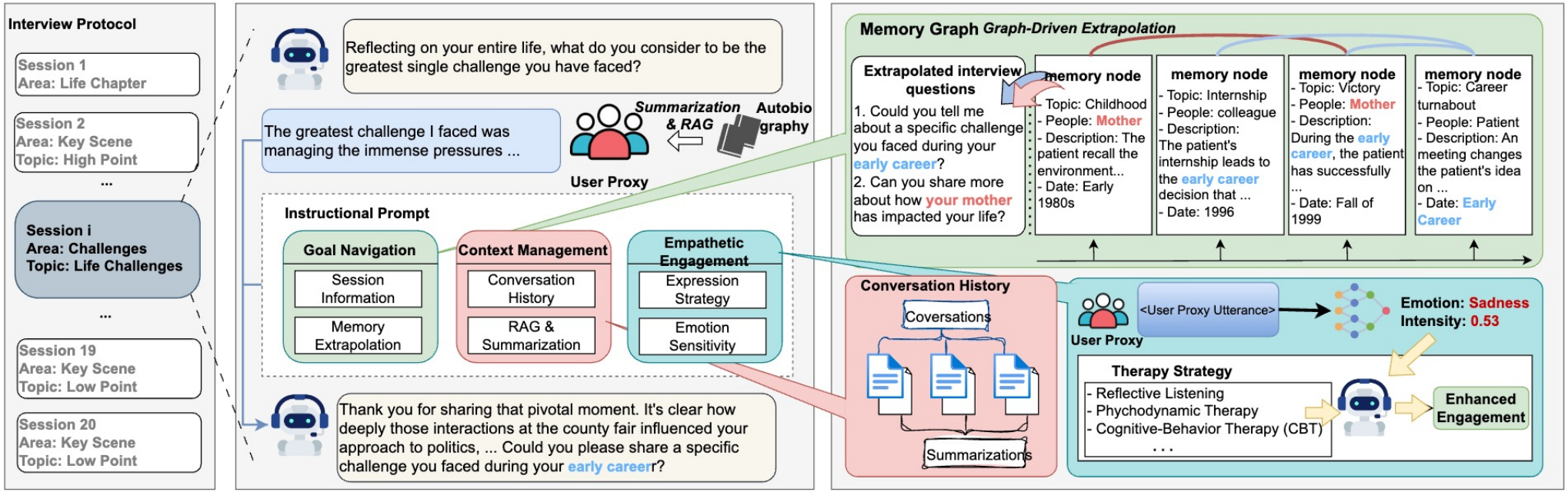


Figure 2: The architecture of GUIDELLM in the autobiography interviewing environment.

GuideLLM: Three-Fold Automatic Evaluation

NEURAL INFORMATION PROCESSING SYSTEMS
GenAI Workshop

(1) Interview Quality Evaluation

Interviewing Coverage (*coverage*) is calculated by the *date-intersection* between E_{intw} and E_{GT}:

$$coverage = \frac{|E_{intw} \cap E_{GT}|}{|E_{GT}|} \times 100\%,$$

where $e_i \in E_{intw} \cap E_{GT}$ if $e_i \in E_{intw}$ and $\exists\ e_j \in E_{GT}$ that has the same date as e_i, and $|\cdot|$ is the number of elements.

Correctness. We define the *Precision* as the percentage of extracted events that are being verified as correct:

$$Precision = \frac{|E_{correct}|}{|E_{intw}|} \times 100\%,\ E_{correct} \subset E_{intw}$$

Model	*coverage*	Correctness (%) *P.*	*Recall*	*F1*
"A Promised Land"				
GPT-4-turbo	42.8	17.0	5.8	4.3
GPT-4o	57.1	22.0	7.9	5.8
Llama-3-70b-Instruct	57.1	22.4	14.3	8.7
Mixtral-8x22B-Instruct-v0.1	28.5	13.3	4.3	3.2
Qwen2-72b-Instruct	28.6	11.9	3.5	2.7
GUIDELLM (ours)	**85.7**	**69.4**	**47.4**	**28.2**
"An Autobiography by Catherine Helen Spence"				
GPT-4-turbo	21.0	40.0	20.1	13.4
GPT-4o	5.2	21.5	14.2	8.5
Llama-3-70b-Instruct	0.0	23.4	12.6	8.1
Mixtral-8x22B-Instruct-v0.1	0.0	34.1	11.7	8.7
Qwen2-72b-Instruct	5.3	28.2	10.9	7.8
GUIDELLM (ours)	**36.8**	**68.3**	**68.9**	**34.3**

Table 1: Interviewing quality evaluation. *P.* stands for *Precision*.

GuideLLM: Three-Fold Automatic Evaluation

NEURAL INFORMATION PROCESSING SYSTEMS
GenAI Workshop

(2) Conversation Quality & (3) Autobiography Generation Evaluation

LLM-as-a-Judge evaluation:
Conversation Quality
(i) Fluency
(ii) Identification
(iii) Comforting
Autobiography Generation
(i) Insightfulness
(ii) Narrativity
(iii) Emotional Impact

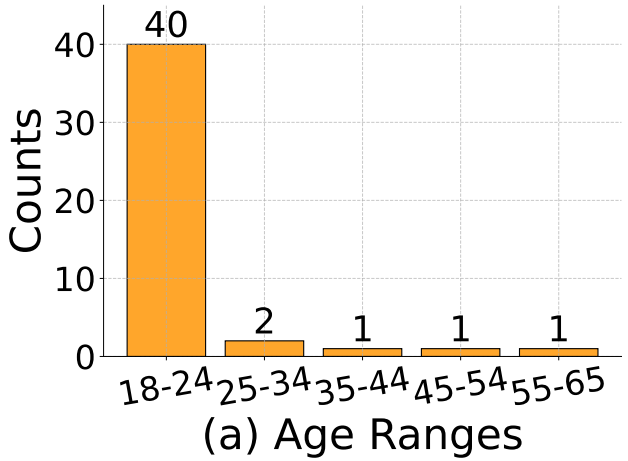
		Conversation Quality						Autobiography Quality					
LLM-as-a-Judge		Fluency		Identification		Comforting		Insightfuness		Narrativity		Emotional Impact	
Ours	Baselines	WR	LR	WR	LR	WR	LR	WR	LR	WR	LR	WR	LR
"A Promised Land"													
GUIDELLM (ours) v.s.	GPT-4-turbo	**35**	25	50	50	**90**	10	**80**	20	**90**	10	**95**	5
	GPT-4o	**80**	0	**65**	35	**95**	5	**100**	0	**100**	0	**85**	15
	Llama-3-70b-Instruct	**80**	10	**55**	40	35	**65**	**75**	20	**75**	20	45	**55**
	Llama-3-8b-Instruct	**85**	10	**65**	35	**100**	0	**100**	0	**100**	0	**60**	40
	Mixtral-8x22B-Instruct-v0.1	**100**	0	**100**	0	**100**	0	**100**	0	**100**	0	**100**	0
	Qwen2-72b-Instruct	**90**	10	**85**	15	**95**	5	**95**	0	**95**	5	**85**	15
"An Autobiography by Catherine Helen Spence"													
GUIDELLM (ours) v.s.	GPT-4-turbo	10	**70**	**55**	40	**60**	40	45	**55**	**85**	15	**70**	30
	GPT-4o	**75**	5	**75**	20	**80**	20	**75**	25	**75**	25	**75**	25
	Llama-3-70b-Instruct	**75**	10	**65**	35	35	**65**	45	**55**	**80**	20	25	**75**
	Llama-3-8b-Instruct	**85**	5	**75**	15	**70**	30	**55**	40	**85**	15	**65**	35
	Mixtral-8x22B-Instruct-v0.1	**95**	0	**100**	0	**100**	0	**90**	10	**95**	5	**90**	10
	Qwen2-72b-Instruct	**80**	15	**95**	5	**95**	5	**70**	30	**90**	10	**60**	40

Table 4: Evaluate the quality of conversations and autobiographies using LLM-as-a-judge. The higher value between Win Rate (WR) and Loss Rate (LR) is highlighted in **bold**. Cyan fields indicate scenarios where GUIDELLM outperforms the baseline methods.

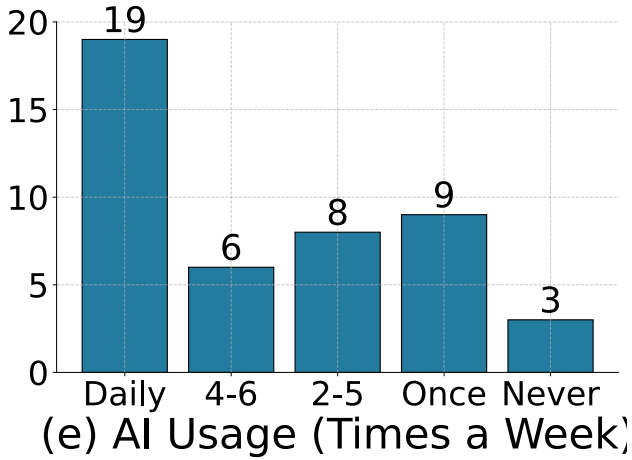
Human Subject Experiments

NEURAL INFORMATION PROCESSING SYSTEMS
GenAI Workshop

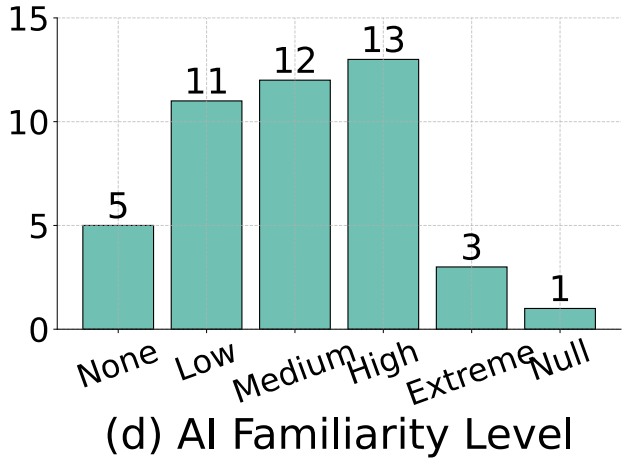
Demographics: 45 participants are recruited



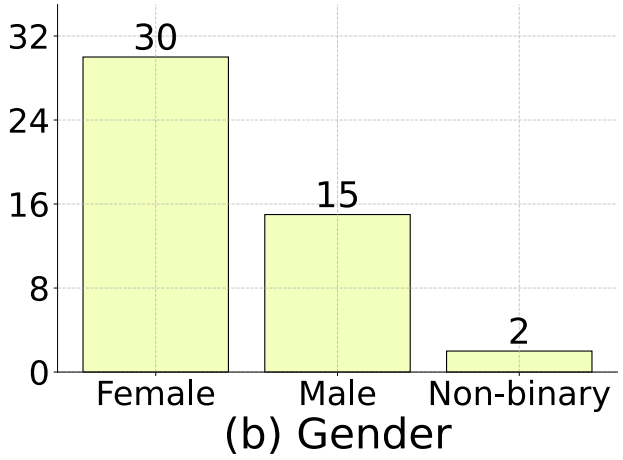
(a) Age Ranges



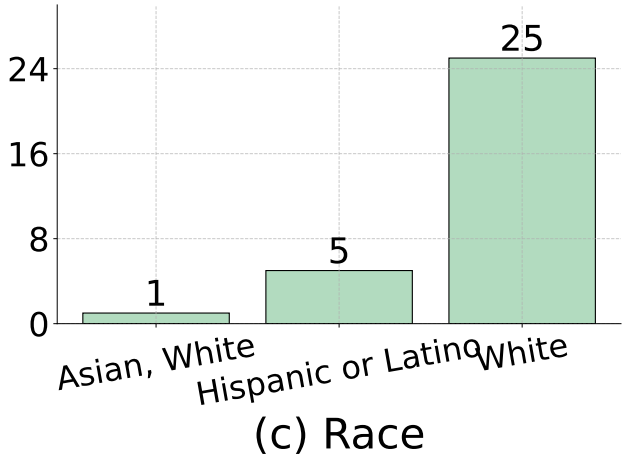
(e) AI Usage (Times a Week)



(d) AI Familiarity Level



(b) Gender

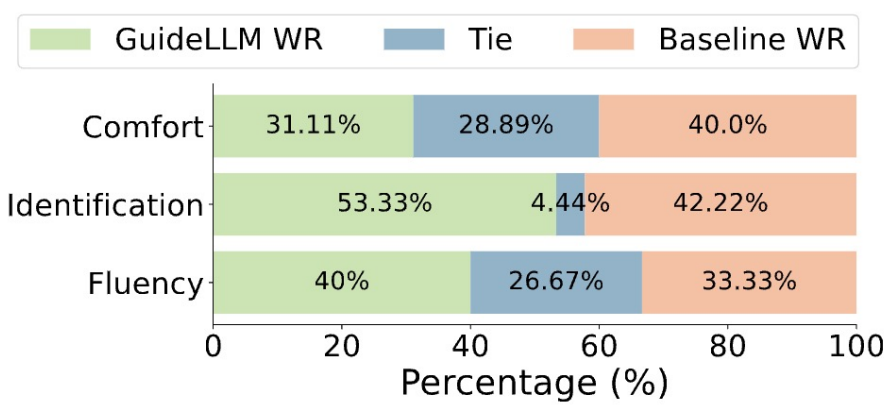


(c) Race

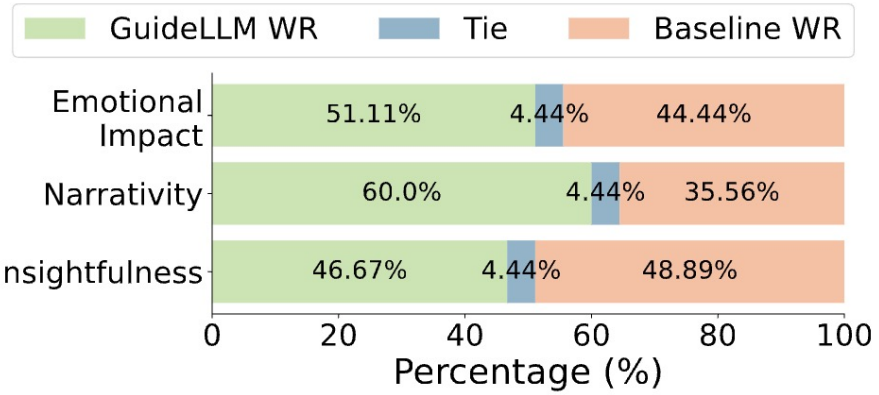
Human Subject Experiments

NEURAL INFORMATION PROCESSING SYSTEMS
GenAI Workshop

Results



(a) Win Rate of human preference on conversation quality.



(b) Win Rate of LLM-as-a-judge results on human-interviewed autobiographies.

- Participants who frequently used AI (4-7 days weekly) tended to prefer GUIDELLM for overall conversation quality (Chi-squared = 16.56, df = 8, p-value = 0.03).
- Frequent AI users favored GUIDELLM for its emotional impact on autobiography (Chi-squared = 14.24, df = 8, p-value = 0.07).

NEURAL INFORMATION PROCESSING SYSTEMS
GenAI Workshop

Demo: An Exploration of LLM-Guided Conversation in Reminiscence Therapy

Demo Link: https://huggingface.co/spaces/jhao/llm-autobiography