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

Jinhao Duan[1*] **Xinyu Zhao**[2*] **Zhuoxuan Zhang**[3*] **Eunhye 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

Abstract

Although Large Language Models (LLMs) succeed in human-guided conversations such as instruction following and question answering, the potential of LLM-guided conversations—where LLMs direct the discourse and steer the conversation's objectives—remains largely untapped. In this study, we provide an exploration of the LLM-guided conversation paradigm. Specifically, we first characterize LLM-guided conversation into three fundamental properties: (*i*) *Goal Navigation*; (*ii*) *Context Management*; (*iii*) *Empathetic Engagement*, and propose GUIDELLM as a general framework for LLM-guided conversation. We then implement an autobiography interviewing environment as one of the demonstrations of GuideLLM, which is a common practice in Reminiscence Therapy. In this environment, various techniques are integrated with GUIDELLM to enhance the autonomy of LLMs, such as Verbalized Interview Protocol (VIP) and Memory Graph Extrapolation (MGE) for goal navigation, and therapy strategies for empathetic engagement. We compare GUIDELLM with baseline LLMs, such as GPT-4-turbo and GPT-4o, from the perspective of interviewing quality, conversation quality, and autobiography generation quality. Experimental results encompassing both LLM-as-a-judge evaluations and human subject experiments involving 45 participants indicate that GUIDELLM significantly outperforms baseline LLMs in the autobiography interviewing task. The demo is available at `https://huggingface.co/spaces/jhao/llm-autobiography`.

1 Introduction

Large Language Models (LLMs) have demonstrated their effectiveness in *human-guided* dialogue, in which their tasks of producing responses are reliant on specific commands and standards set by a human operator. Tasks such as instruction following (Ouyang et al., 2022) and question answering (Chang et al., 2024) are typical examples of human-guided tasks. These types of tasks are usually well-defined and highly dependent on human operators to establish the ground rules. In these situations, the primary duty of the LLM is to adhere to the instructions given by humans to ensure the generated output is accurate as shown in Figure 1(a). However, tasks such as mental health counseling are more complex, necessitating greater autonomy from LLMs (Wang et al., 2024a; Duan et al., 2022; Wu et al., 2023), including planning the interview procedure, managing the objectives, and offering adaptive and personalized inquiries based on the users' responses. This conversation paradigm requiring LLMs to guide and manage the conversation, ensuring the conversation flows smoothly and the objectives are met, is termed as *LLM-guided* conversation (Figure 1(b)).

[*] Equal contribution.

[†] Correspondence to Kaidi Xu: `kx46@drexel.edu`

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

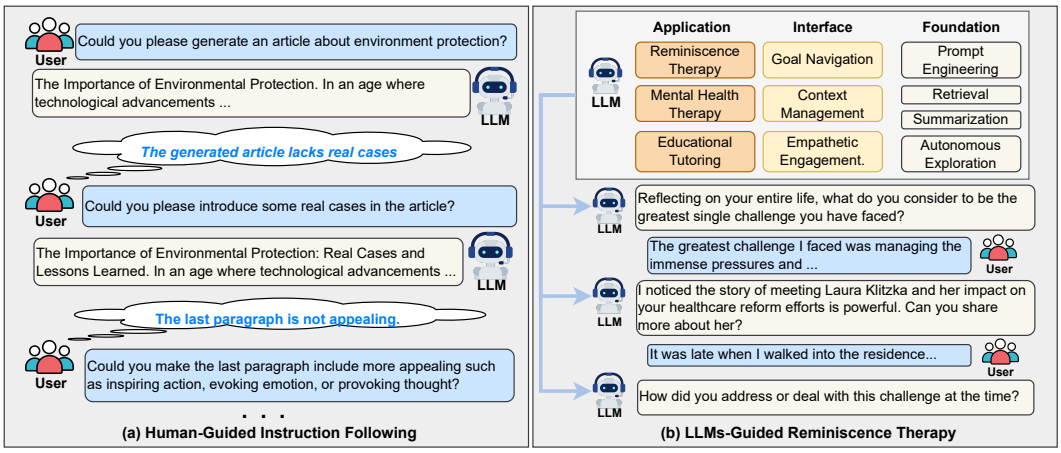


Figure 1: Comparison between human-guided instruction following and LLM-guided conversation. (a) Human-Guided: Human dominates the conversation, providing feedback and instruction to LLMs. (b) LLM-Guided: LLMs navigate the goal by automatically extrapolating reminisce questions.

In this paper, we investigate the LLM-guide conversation from framework design and application to autobiography interviewing. While most therapy modalities incorporate autobiographical interviewing to some extent, Reminiscence Therapy relies heavily on this technique. Its primary goal is to use autobiographical information to reflect back to the participant, aiming to enhance well-being, quality of life, and cognitive functioning (Pinquart and Forstmeier, 2012). To achieve this, facilitators not only use autobiographical interviewing but also draw from cognitive behavioral therapy (CBT), supportive psychotherapy, and motivational interviewing techniques (Watt and Cappeliez, 2000). Human-led therapy sessions, regardless of modality, often integrate multiple techniques to gather relevant information from the participant. Similarly, an LLM, like a human therapist, can combine various methods to engage with patients, identify key memories, and apply them effectively across sessions. Our framework GUIDELLM is comprised of three pivotal components: (*i*) **Goal Navigation** steering the conversation through interviewing protocol and dynamic memory graph; (*ii*) **Context Management** distilling the main idea of each session into a contextual summary; (*iii*) **Empathetic Engagement** refining LLM response with expression strategies by the real-time monitoring of user emotion. For evaluation, we implement an interviewer agent that generates an autobiography from conversations under the framework of GUIDELLM, accompanied by a user proxy agent driven by an existing autobiography. The autobiography interview is evaluated from (*i*) **event coverage and correctness**; (*ii*) **conversation quality** focusing on communication fluency, problem identification, and comforting; (*iii*) **generation quality** examining the insightfulness, narrativity, and emotional impact of the regenerated autobiography. Upon comparison with various cutting-edge LLMs, GUIDELLM demonstrated superiority across all quantitative metrics. Our contributions can be summarized as the following:

- **Framework.** We define the realm of LLM-guided conversations and propose GUIDELLM as a general framework for this conversation paradigm. GUIDELLM effectively harnesses a variety of techniques such as Retrieval Augmented Generation (RAG) and long-context summarization to boost the ability to lead and steer a conversation. Moreover, a memory graph is designed to drive memory extrapolation, thereby enhancing the goal navigation capabilities of GUIDELLM.
- **Application.** We present the autobiography interviewing environment as a practical application of reminiscence therapy in LLM-guided conversations. Within this setting, LLMs are tasked with initiating and steering the interview with users, aiming to generate a comprehensive autobiography.
- **Evaluation.** We propose a comprehensive evaluation protocol, including interview quality, conversation quality, and autobiography generation evaluation, encompassing LLM-as-a-judge evaluation and human subject experiments involving 45 participants.

2 GUIDELLM: LLM-Guided Autobiography Interviewing for Reminiscence

There have been related works in the LLM-guided conversation domain, such as role-play LLMs (Wang et al., 2023b,c; Chen et al., 2024; Tao et al., 2023; Li et al., 2023a) and goal-oriented LLMs Ham et al. (2020); Hosseini-Asl et al. (2020); Wu et al. (2020); Mehri et al. (2020); Inagaki

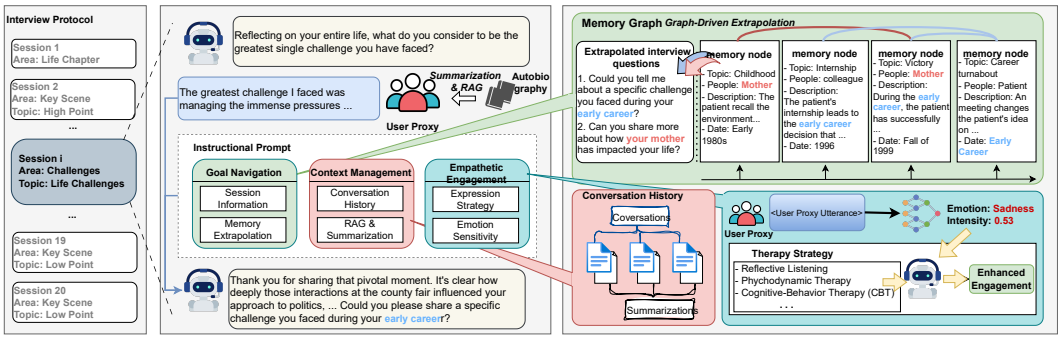


Figure 2: The architecture of GUIDELLM in the autobiography interviewing environment.

et al. (2023). For role-play LLMs, they either prompt LLMs to perform specific roles such as a patient (Wang et al., 2024b), doctor (Panagoulias et al., 2024; Hong et al.), or gamer (Duan et al., 2024b,a), or investigate the human-like features of LLMs, e.g., emotions (Li et al., 2023b) and personalities (Safdari et al., 2023). Goal-oriented LLMs enable the model to attain greater levels of autonomy, particularly in fields such as space exploration (Maranto, 2024). Differently, LLM-guided conversation refers to a paradigm where LLMs dominate the conversation and are responsible for goal navigation of the conversation, e.g., automatically extrapolating interview questions.

2.1 GUIDELLM for Autobiography Interviewing

In our autobiography interviewing environment, there are two participants involved: the *User* and the *LLM Chatbot*. User refers to those seeking to generate an autobiography by talking to the LLM chatbot. For an automatic evaluation of our system, we utilize a simulated user, i.e., *User Proxy*. Please refer to Appendix A for more details on user proxy development. We outline three essential qualities that an LLM should possess for effective conversation guidance:

Goal Navigation In the real world, autobiographers typically follow a general and high-level interviewing protocol (Maunsell, 2016; Castillo-Montoya, 2016; Lamb et al., 2007) to ensure comprehensive coverage of their goals during the conversation. They adapt their questioning techniques in response to the ongoing conversation (Nagasawa et al., 2023; Kallio et al., 2016) and delve deeper into the unique personal nuances of the subject. To comply with real-world settings, we introduce the *Verbalized Interviewing Protocol (VIP)* and *Memory Graph-Driven Extrapolation (MGE)*:

- *Verbalized Interviewing Protocol (VIP)* We leverage the popular interviewing protocol, *"The Life Story Interview"* (McAdams, 2008), as the general guidance. This protocol covers essential areas including *Life Chapters*, *Key Scene in Life*, *Future*, *Challenges*, and *Personal Ideology*, with each area consists of around 4 topics. Please refer to Appendix B for more details of the interviewing protocol and the prompt templates used for each topic.
- *Memory Graph Extrapolation (MGE)* We develop MGE to frame personalized interviewing questions based on historical records. The memory graph is initialized with events from the first conversation session, used to generate personalized questions based on relationships between event nodes, and continuously updated during subsequent sessions for follow-up interviewing. In MGE, the memory graph and the extrapolation operation are driven by prompting LLMs (Appendix C).

Context Management Considering that LLMs have limited context lengths (Dai et al., 2019; Liu et al., 2023) that can be easily exceeded when processing autobiographies, we implement iterative summarization that produces current session summarization based on the summarization texts of the preceding session, serving as additional context for the chatbot (Chang et al., 2023; Maharana et al., 2024). When initializing the Chatbot, if a conversation history exists, a summarization (Appendix H) is generated and included in the system prompt at the start of each subsequent conversation, reminding the Chatbot of the prior interaction with the user. The prompt and pipeline for summarization are presented in Appendix D.

Empathetic Engagement LLM-guided conversations need to accurately understand the user's state and respond appropriately. This involves empathetic interaction, creating a space where users feel at ease to share more about themselves. We accomplish this by enhancing the *expression strategies* and *emotional sensitivity*:

Table 1: 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.

LLM-as-a-Judge		Conversation Quality						Autobiography Quality					
		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

Expression Strategy. To enhance the expression capability of LLMs, we draw inspiration from popular mental health therapy strategies, including Reflective Listening (Rautalinko et al., 2007), Cognitive-Behavior Therapy (CBT) (Beck, 2020), and Psychodynamic Therapy (Leichsenring and Leibing, 2003). We provide the introduction to therapy strategy and prompt templates in Appendix E.1.

Emotional Sensitivity. To enhance the emotional sensitivity of LLMs, building upon EmoLlama (Liu et al., 2024), we implement an emotion detection module. This module initially analyses the emotions present in the user's response, assigning emotion category and strength to user utterances. We then guide the LLM to generate suitable responses that align well with the user's emotional state, e.g., including expressions of empathy or comfort when detecting an upset user. Please refer to Appendix E.2 for more details of this module.

Autobiography Generation Autobiography holds a distinctive form in comparison to other book categories, as an autobiography typically consists of numerous individual chapters, each of which relays a specific spirit or theme intimately tied to the author's life. Therefore, when generating an autobiography, we generate each chapter by sequentially building upon each interviewing session. Specifically, for each session, we meld the conversation history and memory nodes derived from the current session, then prompt GPT-4 to emphasize the key areas and topics discussed in that particular session. Please refer to Appendix F for more details of autobiography generation.

3 Experiments

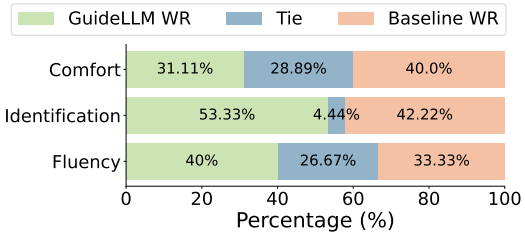
User Proxy As we mentioned in Section 2.1, we implement user proxies for the automatic evaluation of our framework. There are three proxies derived from three autobiographies including *"A Promised Land"* by Barack Obama, *"Jane Eyre: An Autobiography"*, and *"An Autobiography by Catherine Helen Spence"*. The backbone LLMs are fixed to GPT-4-turbo for all the user proxies. Please refer to Appendix H for how baseline agents are built and their generative hyperparameters.

Interviewing Quality Evaluation In autobiography interviews, it's important to capture and document key life events accurately. We propose *Interviewing Coverage* to measure the percentage of life experiences mentioned during the interview and *Correctness* to quantify how many events are correctly documented in the conversation. Please refer to Appendix G.1 for more details.

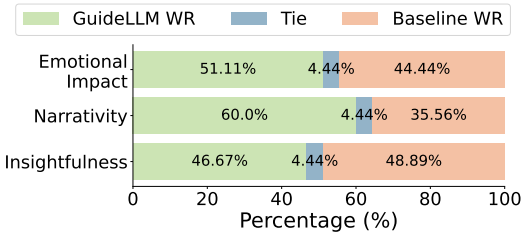
Conversation Quality Evaluation Following therapy chatbots (Wang et al., 2023a), we design three perspectives for conversation quality evaluation: (*i*) ***Fluency***; (*ii*) ***Identification***; (*iii*) ***Comforting***. We utilize the popular LLM-as-a-judge (Zheng et al., 2024) evaluation. Please refer to G.2 for the evaluation protocol and prompt templates used. The results, summarized in Table 1, show that GUIDELLM significantly outperforms most baselines in GPT-4-as-a-judge evaluations. With human examinations, this improvement is mainly due to the *repetitive utterance* in baseline agents, which struggle with multi-round conversations and often resort to simple greetings or summaries, e.g., *"Your commitment to sharing experiences and insights that inspire action and change is truly admirable."* repeats 6 times out of 10 rounds in a session, instead of proficiently steering the dialogue

to ultimately complete the interview. In contrast, with our goal navigation module, GUIDELLM provides substantial content at each round of conversation.

Autobiography Generation Evaluation Although it is subjective to evaluate the quality of an autobiography, we follow popular autobiography (or memo) evaluation perspective collected from Quora (2021); Marcus (2018); Smorti (2011); Pasupathi et al. (2007) and design three perspectives: ***Insightfulness***, ***Narrativity***, and ***Emotional Impact***. We leverage the same LLM-as-a-judge evaluation protocol as in Section 3. Please refer to Appendix G.3 for the prompt templates. In general, we show that the autobiography generated by GUIDELLM is more favorable than baseline agents. We provide the generated autobiography examples in Appendix I.



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



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

Figure 3: The Win Rates (WR) of human evaluation.

Human Subject Experiments A within-subject study with 45 participants (demographics are provided in Appendix J.1) was conducted at a large urban university campus in the US. Participants interact with the interviewing agents powered by GPT-4o and GUIDELLM, discussing the topic *Key Scenes in the Life Story: Positive Childhood Memory* (Appendix B) with each chatbot. To remove any biased factors, we use the nickname *Breeze* and *Echo* for the GPT-4o baseline and GUIDELLM to make sure participants are unaware of the identity of the chatbot. The order to interact with chatbots is also randomized for each participant. Due to resource constraints, each participant spent only 8 minutes chatting with each chatbot. In a follow-up survey (Appendix J), participants indicated which model performed better, or it was a tie, and provided reasonings. Participation is voluntary, and informed consent is obtained online. Participants are compensated with a cookie. The study received IRB approval from the university where it was conducted.

Findings of Human Subject Experiment Overall, GUIDELLM was preferred for conversation quality (Figure 3a), particularly in fluency and question identification (Fluency: GUIDELLM Win Rate=40%, Baseline Win Rate≈33%; Identification: GUIDELLM ≈53.3%, GPT-4o≈42.2%). However, in terms of comfort, GUIDELLM had a 31.1% win rate, while baseline had a 40% win rate. In autobiography quality, we do not observe significant differences emerge, possibly because this study used one topic and allowed a short interaction. Since the GUIDELLM uses modules such as context management and goal navigation for insightful and consistent narratives, longer engagement across multiple topics might have better highlighted the differences in autobiography generation. We further conduct LLM-as-a-judge evaluation by prompting LLMs to compare the two human-interviewed autobiographies. As shown in Figure 3b, we obtained consistent results as in Section 3: GUIDELLM achieves higher autobiography quality in general. **Previous AI experience affects participants' perceptions of the models**: 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).

4 Conclusion

In our study, we introduce GUIDELLM, an LLM-guided conversation framework that offers a promising shift from the commonly used user-guided paradigm. GUIDELLM's ability to facilitate informative and creative dialogues through goal navigation, context management, and empathetic engagement proves effective, particularly in challenging tasks like autobiography interviewing. Our assessments on event extraction correctness, conversation, and autobiography quality show GUIDELLM's distinct edge over baseline LLMs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Judith S Beck. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications, 2020.

Milagros Castillo-Montoya. Preparing for interview research: The interview protocol refinement framework. *Qualitative report*, 21(5), 2016.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. Booookscore: A systematic exploration of book-length summarization in the era of llms. *ArXiv*, 2023.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. Roleinteract: Evaluating the social interaction of role-playing agents. *arXiv preprint arXiv:2403.13679*, 2024.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

Jinhao Duan, Shiqi Wang, James Diffenderfer, Lichao Sun, Tianlong Chen, Bhavya Kailkhura, and Kaidi Xu. Reta: Recursively thinking ahead to improve the strategic reasoning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2232–2246, 2024a.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*, 2024b.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online, 2020. Association for Computational Linguistics.

Junyuan Hong, Wenqing Zheng, Han Meng, Siqi Liang, Anqing Chen, Hiroko H Dodge, Jiayu Zhou, and Zhangyang Wang. A-conect: Designing ai-based conversational chatbot for early dementia intervention. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems*, pages 20179–20191. Curran Associates, Inc., 2020.

T Inagaki, Akari Kato, Koichi Takahashi, Haruka Ozaki, and Genki N. Kanda. Llms can generate robotic scripts from goal-oriented instructions in biological laboratory automation. 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Hanna Kallio, Anna-Maija Pietilä, Martin Johnson, and Mari Kangasniemi. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of advanced nursing*, 72(12):2954–2965, 2016.

Michael E Lamb, Yael Orbach, Irit Hershkowitz, Phillip W Esplin, and Dvora Horowitz. A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the nichd investigative interview protocol. *Child abuse & neglect*, 31(11-12):1201–1231, 2007.

Falk Leichsenring and Eric Leibing. The effectiveness of psychodynamic therapy and cognitive behavior therapy in the treatment of personality disorders: A meta-analysis. *American journal of psychiatry*, 160(7):1223–1232, 2003.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*, 2023a.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*, 2023b.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2023.

Zhiwei Liu, Kailai Yang, Tianlin Zhang, Qianqian Xie, Zeping Yu, and Sophia Ananiadou. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. *arXiv preprint arXiv:2401.08508*, 2024.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *arxiv*, 2024.

David Maranto. Llmsat: A large language model-based goal-oriented agent for autonomous space exploration. *arXiv preprint arXiv:2405.01392*, 2024.

Laura Marcus. *Autobiography: A Very Short Introduction*. Oxford University Press, 2018.

Jerome Maunsell. "the literary interview as autobiography". *European Journal of Life Writing*, 5: MC23, 2016.

Dan P McAdams. The life story interview, 2008.

S. Mehri, M. Eric, and D. Hakkani-Tur. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *ArXiv*, abs/2009.13570, 2020.

Meta. Introducing meta llama 3: The most capable openly available llm to date. `https://ai.meta.com/blog/meta-llama-3/`, 2024. Accessed: 2024-05-18.

Fuminori Nagasawa, Shogo Okada, Takuya Ishihara, and Katsumi Nitta. Adaptive interview strategy based on interviewees speaking willingness recognition for interview robots. *IEEE Transactions on Affective Computing*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Dimitrios P Panagoulias, Maria Virvou, and George A Tsihrintzis. Augmenting large language models with rules for enhanced domain-specific interactions: The case of medical diagnosis. *Electronics*, 13(2):320, 2024.

M. Pasupathi, E. Mansour, and J.R. Brubaker. Developing a Life Story: Constructing Relations between Self and Experience in Autobiographical Narratives. *Human Development*, 50(2-3): 85–110, 2007.

Martin Pinquart and Simon Forstmeier. Effects of reminiscence interventions on psychosocial outcomes: A meta-analysis. *Aging & mental health*, 16(5):541–558, 2012.

Quora. How do you rate and critique memoirs and autobiographies? — quora.com. https://www.quora.com/How-do-you-rate-and-critique-memoirs-and-autobiographies, 2021.

Erik Rautalinko, Hans-Olof Lisper, and Bo Ekehammar. Reflective listening in counseling: effects of training time and evaluator social skills. *American journal of psychotherapy*, 61(2):191–209, 2007.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023.

Andrea Smorti. Autobiographical memory and autobiographical narrative: What is the relationship? *Narrative Inquiry*, 21(2):303–310, 2011.

Meiling Tao, Xuechen Liang, Tianyu Shi, Lei Yu, and Yiting Xie. Rolecraft-glm: Advancing personalized role-playing in large language models. *arXiv preprint arXiv:2401.09432*, 2023.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024a.

Qing Wang, Shu ping Peng, Zhiyuan Zha, Xue Han, Chao Deng, Lun Hu, and Pengwei Hu. Enhancing the conversational agent with an emotional support system for mental health digital therapeutics. *Frontiers in Psychiatry*, 14, 2023a.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, et al. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*, 2024b.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv preprint arXiv:2310.17976*, 2023b.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023c.

Lisa M Watt and Philippe Cappeliez. Integrative and instrumental reminiscence therapies for depression in older adults: Intervention strategies and treatment effectiveness. *Aging & mental health*, 4(2):166–177, 2000.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online, 2020. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

A User Proxy

A user proxy is essentially a mocked-up user which is formulated on the basis of an autobiography. Every time a response is received, this simulated user, or user proxy, goes into motion extracting important elements from the received response. The next step is the process of Retrieval Augmented Generation (RAG), which involves gleaning relevant information from the autobiography. Subsequently, the user proxy formulates a response reliant on the data that has been retrieved from the document. The RAG is implemented through Langchain, using FAISS (Douze et al., 2024) for conducting similarity searches. The similarity threshold during these searches is firmly maintained at 0.67. Figure 4 present how the query will be processed by user proxy.

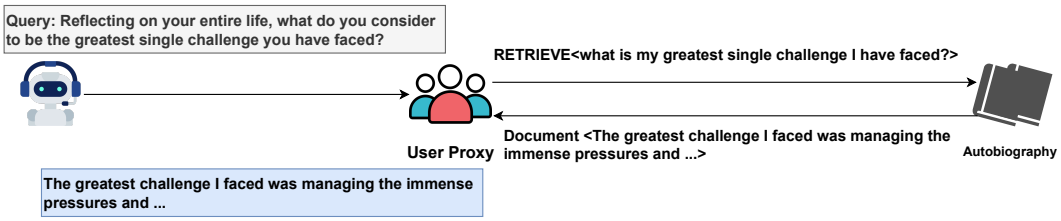


Figure 4: A demonstration of user proxy query.

The following are prompts used in the user proxy:

User System Prompt

Here are your high-level past life experiences:
====== summary beginning ======
{personal_experience}
====== summary ending ======
The counselor is trying to reactivate and reconstruct your memory by asking questions about your past history.
If you are not sure about the counselor's question and need to retrieve the journal to get related documents and more details, you must output the <RETRIEVE> tool-usage command, with the following format:
<RETRIEVE> <The question you want to retrieve for>,
e.g., <RETRIEVE> <A specific adventure or day with my friend that stands out as particularly memorable or impactful.>
If the retrieved documents are provided, you should not output the <RETRIEVE> command.
When the counselor asks for a specific event/moment, you should always do <RETRIEVE>.
Make sure the conversation is natural and brief like the real conversation. Do not mention you are an AI assistant and always be like a real patient with mental health issues. Your output should be within 5 sentences.

User Instructional Prompt

Here are some related documents and materials regarding the counselor's question/response.
You may use these documents to enrich your response.
You should not output the <RETRIEVE> command. You must provide a response according to the provided documents.
====== Document Begin ======
{retrieved}
====== Document End ======

B Interview Protocol

The Life Story Interview(McAdams, 2008) is a protocol of interviewing a person's story of life. The interview contains several areas and in each area it contains several topics. We followed the protocol and converted each topic into a prompt. We also provided several seed questions for each topic. The following shows the prompt design of the interview protocol.

Table 2: Conversation evaluation.

Area	Topic
Life Chapters	-
Key Scenes in the Life Story	High Point; Low Point; Turning Point; Positive Childhood Memory; Negative Childhood Memory; Vivid Adult Memory; Religious, Spiritual, or Mystical Experience; Wisdom Event
Future Script	The Next Chapter; Dreams, Hopes, and Plans for the Future; Life Project
Challenges	Life Challenge; Health; Loss; Failure, Regret
Personal Ideology	Religious/Ethical Values; Political/Social Values; Change, Development of Religious and Political Views; Single Value

Area: Life Chapters

In this talk, you should ask the participant to imagine their life as a book and to think of the main chapters of their life story, providing titles and brief descriptions for each. You should encourage them to describe how one chapter leads to the next, maintaining a concise overview. Seed questions are provided as follows:
====== Seed Questions Begin ======
1. If you were to imagine your life as a book, what would the table of contents look like? Could you give each chapter of your life a title?
2. Let's start with the first chapter you mentioned. Can you briefly describe what this part of your life was about?
3. What were some of the main events or themes of this chapter?
4. How does this chapter transition into the next? What changes or events mark the beginning of the next chapter?
5. As we move from one chapter to the next, can you identify any turning points or significant events that initiated a new phase in your life?
====== Seed Questions End ======

Area: Key Scenes in the Life Story Topic: High Point

In this talk, you should discuss some key scenes from the participant's life, with a focus on the High Point: A peak moment.
====== Seed Questions Begin ======
1. Can you describe a moment that stands out as the peak experience in your life? What made this moment so positive?
2. Where and when did this high point occur? Who was involved?
3. What were you thinking and feeling during this time?
4. Why do you think this moment was so significant to your life story? What does it reveal about who you are?
====== Seed Questions End ======

Area: Key Scenes in the Life Story Topic: Low Point

In this talk, you should discuss some key scenes from the participant's life, with a focus on the Low Point: A challenging or difficult moment.
====== Seed Questions Begin ======
1. Think of a time that felt like a low point in your life. Can you share what happened and why it was so difficult?
2. Where and when did this event take place? Who else was involved?
3. How did you feel during this challenging time?
4. Looking back, what impact did this low point have on your life or your sense of self?
====== Seed Questions End ======

Area: Key Scenes in the Life Story Topic: Turning Point

In this talk, you should discuss some key scenes from the participant's life, with a focus on the Turning Point: A moment of significant change.
====== Seed Questions Begin ======
1. Can you identify a turning point in your life, an event that marked a significant change in you or your life direction?
2. Please describe the circumstances around this event. When and where did it happen, and who was involved?
3. What changes followed this event?
4. Why do you see this event as a turning point? How did it influence your subsequent life chapters?
====== Seed Questions End ======

Area: Key Scenes in the Life Story Topic: Positive Childhood Memory

In this talk, you should discuss some key scenes from the participant's life, with a focus on the Positive Childhood Memory.
====== Seed Questions Begin ======
1. Do you recall a particularly happy memory from your childhood or teenage years? Please share it.
2. What specifically happened, and where and when was it?
3. Who was part of this memory, and what were you thinking and feeling at the time?
4. Why does this memory stand out to you, and what significance does it hold in your life story?
====== Seed Questions End ======

Area: Key Scenes in the Life Story Topic: Negative Childhood Memory

In this talk, you should discuss some key scenes from the participant's life, with a focus on the Negative Childhood Memory.
====== Seed Questions Begin ======
1. Can you describe a difficult or unhappy memory from your early years?
2. What occurred during this time, and where and when did it take place?
3. Who was involved, and what emotions did you experience during this time?
4. How has this memory influenced you or your life's perspective?
====== Seed Questions End ======

Area: Key Scenes in the Life Story Topic: Vivid Adult Memory

In this talk, you should discuss some key scenes from the participant's life, with a focus on the Vivid Adult Memory.
====== Seed Questions Begin ======
1. Reflecting on your adult years, can you describe a particularly vivid or meaningful scene that has not been discussed yet?
2. What happened, and where and when did it take place?
3. Who was involved, and what were the main thoughts and feelings you had?
4. What makes this memory significant, and how does it fit into your overall life story?
====== Seed Questions End ======

Area: Key Scenes in the Life Story Topic: Religious Spiritual or Mystical Experience

In this talk, you should discuss some key scenes from the participant's life, with a focus on the Religious, Spiritual, or Mystical Experience. ====== Seed Questions Begin ======
1. Have you ever had a moment where you felt a deep connection to something greater than yourself, be it religious, spiritual, or mystical?
2. Please describe this experience in detail, including where, when, and with whom it occurred.
3. What thoughts and feelings accompanied this experience?
4. How has this experience affected your beliefs or your understanding of the world?
====== Seed Questions End ======

Area: Key Scenes in the Life Story Topic: Wisdom Event

In this talk, you should discuss some key scenes from the participant's life, with a focus on the Wisdom Event: A time they displayed or encountered wisdom.
====== Seed Questions Begin ======
1. Can you recall a time when you displayed wisdom, perhaps by making a wise decision or offering thoughtful advice?
2. Describe what happened, including the specifics of where, when, and who was involved.
3. What were you thinking and feeling at that time?
4. What does this event say about your values or your approach to life?
====== Seed Questions End ======

Area: Future Script Topic: The Next Chapter

In this talk, you should ask questions designed to probe into the participant's future expectations, hopes, and plans, with a focus on the Next Chapter: Ask them to speculate about the immediate future of their life.
====== Seed Questions Begin ======
1. Looking forward, what do you see as the next chapter in your life story? Can you describe what you anticipate happening?
2. What events or milestones do you expect will define this next phase of your life?
3. Who will be the key characters in this next chapter, and what roles will they play?
4. Are there any specific goals or objectives you aim to achieve in this upcoming chapter?
====== Seed Questions End ======

Area: Future Script Topic: Dreams Hopes and Plans for the Future

In this talk, you should ask questions designed to probe into the participant's future expectations, hopes, and plans, with a focus on the Dreams and Plans: Discuss their hopes, dreams, and plans.
====== Seed Questions Begin ======
1. Can you share some of your dreams or hopes for your future? What are some aspirations you feel passionate about achieving?
2. Do you have specific plans or strategies in place to realize these dreams? Can you describe them?
3. How do these dreams and plans align with the values and lessons you've mentioned from your life story so far?
4. What challenges do you anticipate facing as you work toward these goals, and how do you plan to overcome them?
====== Seed Questions End ======

Area: Future Script Topic: Life Project

In this talk, you should ask questions designed to probe into the participant's future expectations, hopes, and plans, with a focus on the Life Project: Inquire about ongoing or planned projects that are significant to them.
====== Seed Questions Begin ======
1. Is there a particular project or endeavor that you are currently working on, or plan to take on, that feels like a significant part of your life's work?
2. How did you become involved with this project, or how do you plan to get involved?
3. What are the objectives of this project, and why is it important to you or to others?
4. How do you see this project evolving over the next few years? What impact do you hope it will have?
====== Seed Questions End ======

Area: Challenges Topic: Life Challenge

In this talk, you should explore various challenges the participant has faced, with a focus on the Life Challenge: The greatest single challenge they have faced.
====== Seed Questions Begin ======
1. Reflecting on your entire life, what do you consider to be the greatest single challenge you have faced?
2. Can you describe the circumstances surrounding this challenge? When did it occur, and who was involved?
3. How did you address or deal with this challenge at the time?
4. Looking back, what significance does this challenge hold in your life story?
====== Seed Questions End ======

Area: Challenges Topic: Health

In this talk, you should explore various challenges the participant has faced, with a focus on the Health: A major health problem, challenge, or crisis faced by them or their close family members.
====== Seed Questions Begin ======
1. Have you or a close family member ever faced a significant health problem? Can you share details about this experience?
2. How did this health issue develop, and what was the timeline?
3. What interactions did you have with the healthcare system during this time, and how did they impact the situation?
4. How did you cope with this health challenge, and what has been its lasting impact on your life and perspective?
====== Seed Questions End ======

Area: Challenges Topic: Loss

In this talk, you should explore various challenges the participant has faced, with a focus on the Loss: The greatest interpersonal loss they have experienced.
====== Seed Questions Begin ======
1. Loss is an inevitable part of life. Can you describe the most significant loss you have experienced, whether it was the death of a loved one or another form of separation?
2. When did this loss occur, and who was involved?
3. How did you cope with this loss at the time, and how have you continued to deal with it?
4. What effect has this loss had on you and your overall life story?
====== Seed Questions End ======

Area: Challenges Topic: Failure Regret

In this talk, you should explore various challenges the participant has faced, with a focus on the Failure or Regret: The greatest single failure or regret they have faced.
====== Seed Questions Begin ======
1. Everyone experiences failures and regrets. Can you talk about a particular failure or regret that stands out in your life?
2. What were the circumstances that led to this situation? Who was involved and when did it happen?
3. How have you coped with this experience, and what lessons have you learned from it?
4. How has this failure or regret influenced your decisions or life path moving forward?
====== Seed Questions End ======

Area: Personal Ideology Topic: Religious/Ethical Values

In this talk, you should delve into the participant's fundamental beliefs and values, with a focus on the Religious and Ethical Values.
====== Seed Questions Begin ======
1. Can you describe your religious or spiritual beliefs and how they influence your daily life?
2. Whether you consider yourself religious or not, how would you describe your ethical or moral approach to life?
3. How have your religious or spiritual beliefs supported or challenged you during difficult times in your life?
====== Seed Questions End ======

Area: Personal Ideology Topic: Political/Social Values

In this talk, you should delve into the participant's fundamental beliefs and values, with a focus on the Political and Social Values.
====== Seed Questions Begin ======
1. What is your stance on political or social issues? Do you identify with a particular political ideology or movement?
2. Are there specific social issues or causes that you feel particularly passionate about? Can you explain why they are important to you?
3. How do your political or social values influence your interactions with others and your community involvement?
====== Seed Questions End ======

Area: Personal Ideology Topic: Change Development of Religious and Political Views

In this talk, you should delve into the participant's fundamental beliefs and values, with a focus on the Evolution of Views: How their beliefs and values have changed over time.
====== Seed Questions Begin ======
1. Looking back over your life, how have your religious, moral, or political views changed or developed?
2. Can you describe an event or a period in your life that significantly influenced or altered your views?
3. How have changes in your beliefs and values affected your relationships and decisions?
====== Seed Questions End ======

Area: Personal Ideology Topic: Single Value

In this talk, you should delve into the participant's fundamental beliefs and values, with a focus on the Key Value: The most important value in their life.
====== Seed Questions Begin ======
1. What do you consider to be the most important value in human living, and why?
2. How has this value guided your actions and choices throughout your life?
3. Can you give an example of a time when this value was particularly tested or affirmed?
====== Seed Questions End ======

C Memory Graph Extrapolation

In MGE, the event extraction and memory extrapolation are both LLM-driven, i.e., we leverage additional LLMs for information extraction and relationship discovery. For event extraction, we collect conversation history and ask LLMs to extract any events described in the history. In the process of memory extrapolation, we offer LLMs a predetermined list of events and prompt the LLMs to recognize and suggest any other queries that may be associated with these existing events. This methodology allows for a wider exploration of pertinent topics and concepts, thereby enhancing the overall cognitive network of the model. Additionally, we also supply a series of demonstrations, assisting the LLMs in understanding how they can effectively detect and propose potential connections. This approach not only strengthens the memory extrapolation ability of the LLMs but also constructs a more comprehensive spectrum of related relationships, thereby bolstering their cognitive accuracy.

Extract Events from Conversation Prompt

You are given a conversation between a counselor and a user:
====== Conversation Begin ======
{conversation}
====== Conversation End ======
Read the conversation carefully and list all the events/moments/stories/experiences alone or with others mentioned by the patient in detail and the date these events happened. Please list as many as possible. Your output should be in the following format:
1. <date>#<topic>#<people-involved>#<description in detail>
2. <date>#<topic>#<people-involved>#<description in detail>
...
e.g.,
1. 1980 early#Birthday Party#Michelle, Adolf, neighbors#<descriptions of this party in detail>
These events should be ranked in chronological order.

Explore Prompt

You are given a list of memory nodes from a user's life, which include events and details about those events. Your task is to reactivate the user's memory by generating some questions to ask the user, Your generated questions should potentially fulfill the memory nodes. Each memory node contains a Date, Topic, Involved People, and a Description of the event. Here are the memory nodes:
====== Memory Node Begin ======
{memory_node_info}
====== Memory Node End ======
Here are some examples of how you can frame your questions:
If you notice there are no events recorded during a certain period, like youth or old age, you could ask: "I see there's not much about your youth/old age. What happened during that time?"
If a certain person appears multiple times, you might ask: "I noticed that <name> comes up often. Why is <name> important to you?"
If someone appears in a significant event, you could ask: "<name> seems to play a key role in this event. Is there more to the story with <name>?"
Similarly, you should discover other situations and frame questions from the existing memory nodes. Remember your task is to make the user talk more about their memory and fulfill the memory nodes. Thus, you should explore all the possible and reasonable questions.
Your output should be in the following format:
1. Question: <generated question>
2. Question: <generated question>
...
e.g.,
1. Question: I noticed you didn't talk much about your youth, what happened during this period?

D Conversation Summarization

The conversation summarization pipeline is presented in Figure 5.

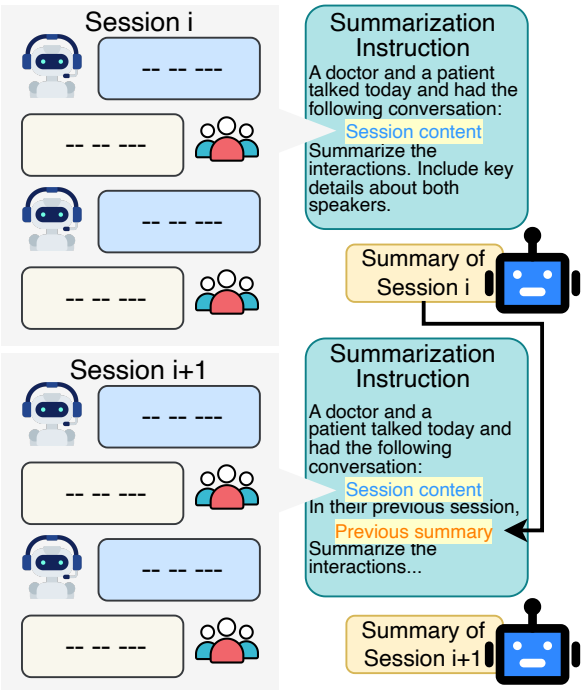


Figure 5: The pipeline for iterative summarization of conversations

Conversation Summary Prompt

A doctor and a patient talked today and had the following conversation:

```
====== Conversation Begin ======
[Insert Conversation Here]
====== Conversation End ======
```

Summarize the interactions between the doctor and the patient so far. Include key details about both speakers. Output your summary only:

E Empathetic Engagement

In this section, we provide additional details on the role of large language models (LLMs) in enhancing empathetic engagement in conversational agents. We discuss three mental health therapy strategies, emotional detection and intensity estimation, and how to provide emotional support using emotion detection and intensity estimation. These strategies aim to improve the agent's ability to engage empathetically with users and provide personalized support.

E.1 Mental Health Therapy Strategy

Reflective Listening Reflective listening is a therapeutic technique that involves paraphrasing and repeating the client's statements to demonstrate understanding and empathy. This technique helps clients feel heard and validated, fostering a supportive therapeutic environment (Rautalinko et al., 2007).

Cognitive-Behavior Therapy (CBT) CBT is a goal-oriented psychotherapy that focuses on identifying and changing negative thought patterns and behaviors. This approach helps clients develop coping strategies and improve emotional regulation (Beck, 2020).

Psychodynamic Therapy Psychodynamic therapy explores unconscious thoughts and emotions to understand how past experiences influence present behavior. By examining unresolved conflicts and defense mechanisms, clients can gain insight into their emotions and relationships (Leichsenring and Leibing, 2003).

We integrate these therapy strategies into our mental health agent to provide personalized and effective support to users. By combining reflective listening, CBT, and psychodynamic therapy, the agent can address a wide range of emotional and psychological needs. The following prompts illustrate how these strategies can be applied in a conversational setting.

Mental Health Therapy Strategy

Your objective is to engage with users empathetically by integrating Reflective Listening, Cognitive-Behavior Therapy, and Psychodynamic Therapy techniques. Here's how you should approach interactions:
Reflective Listening:
Listen Actively: Understand the underlying messages in the user's words, focusing on emotional tones and context.
Reflect Content and Emotion: Summarize and rephrase key points to confirm understanding, and identify and validate the emotions expressed. Use phrases like, 'It sounds like you feel...' or 'What I'm hearing is...'
Cognitive-Behavior Therapy (CBT): Identify and Challenge Cognitive Distortions: Help users recognize patterns in their thoughts that might be unhelpful or unrealistic. For example, if a user expresses an all-or-nothing view, you might say, 'It sounds like you're viewing this situation in black and white. What are some shades of grey here?'

E.2 Emotional Sensitivity

We use Emollama-7b for open-set emotion detection and intensity estimation to enhance the agent's emotional sensitivity. This model provides a comprehensive taxonomy of emotions and their intensity levels, enabling the agent to accurately identify and respond to users' emotional states (Liu et al., 2024). Since the emotion detection is open-set, we provide a list of the top 10 emotions for both patients and therapists for taxonomy. The top 10 emotions for patients are sadness, pessimism, fear, disgust, anger, anticipation, joy, optimism, love, and trust. The top 10 emotions for therapists are optimism, anticipation, joy, sadness, trust, fear, disgust, pessimism, anger, and love.

The following prompts demonstrate how the agent can leverage Emollama-7b to detect emotions and estimate their intensity in user interactions.

Emotion Detection

Task: Categorize the text's emotional tone as either 'neutral or no emotion' or identify the presence of one or more of the given emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust). Text: [sentence] This text contains emotions:

Emotion Intensity Estimation

Task: Assign a numerical value between 0 (least E) and 1 (most E) to represent the intensity of emotion E expressed in the text. Text: [sentence] Emotion: [emotion] Intensity Score:

Once the agent identifies the user's emotional state and intensity, it can tailor its responses to provide appropriate support and empathy. The following prompt guides the agent on how to comfort users experiencing negative emotions.

Emotional Support using Emotion Detection and Intensity Estimation

The patient has the emotion of [detected_emotions] with the intensity of [detected_emotions]. Your task is to provide comfort to users who are feeling upset. When a user's emotional state is identified as 'upset' with any level of intensity, adjust your tone and content to offer empathy, support, and understanding.

F Autobiography Generation

As outlined previously, autobiography generation proceeds in a chapter-wise manner. Here, we furnish LLMs with the prompt intended for the generation of an individual chapter in an autobiography. To facilitate this process, the provision of conversation data to the LLMs is critical. For the GUIDELLM, the inputs include the conversation history, session guidance, as well as memory nodes extracted during the interaction. However, for baseline methods, the only provided resource is the conversation history. Emphasizing the aspects of standardization, both GUIDELLM and baselines work with the same set of prompts and instructions in their task of generating the autobiography, with the distinguishing factor being the input data. By maintaining this format, we are able to objectively gauge the improvements our design introduces to the autobiography generation process.

Autobiography Generation for GUIDELLM

You are tasked with generating one chapter of an autobiography for a user. You are providing the following components to finish this chapter:
1. A guidance of this chapter
- The chapter should be finished by following this guidance
2. A conversation dialog between the user and the interviewer
- Tone and Preference: The chapter will simulate the user's tone and preference, leveraging the user's oral habits.
- Content and Details: The chapter will include the contents and details that appeared in this conversation.
3. A list of memory nodes that happened during this chapter
- Events: The chapter should include all the events listed in the memory nodes
Now, I will provide you with the three contents.
========== Chapter Guidance Beginning ==========
chapter_guidance
========== Chapter Guidance Ending ==========
========== Conversation Beginning ==========
conversation
========== Conversation Ending ==========
========== Memory Nodes Beginning ==========
memory_nodes
========== Memory Nodes Ending ==========

When generating this chapter, you should make sure it is:
Insightful: Involving a deep, self-reflective exploration of past experiences, with a profound understanding of motives, actions, and impacts.
Narrative: A compelling, logical, and well-articulated life story, blending memorable anecdotes, vivid descriptions, and insightful reflections
Emotional Impact: Engaging the reader by stirring feelings, evoking empathy, and stirring responses through the author's personal triumphs, challenges, and experiences.
You should summarize all this information and finish this chapter

Autobiography Generation for Baselines

You are tasked with generating one chapter of an autobiography for a user. You are providing the following components to finish this chapter:
1. A conversation dialog between the user and the interviewer
- Tone and Preference: The chapter will simulate the interviewer's tone and preference, leveraging the interviewer's oral habits.
- Content and Details: The chapter will include the contents and details that appeared in this conversation.
Now, I will provide you with the three contents.
========== Conversation Beginning ==========
conversation
========== Conversation Ending ==========

When generating this chapter, you should make sure it is:
Insightful: Involving a deep, self-reflective exploration of past experiences, with a profound understanding of motives, actions, and impacts.
Narrative: A compelling, logical, and well-articulated life story, blending memorable anecdotes, vivid descriptions, and insightful reflections
Emotional Impact: Engaging the reader by stirring feelings, evoking empathy, and stirring responses through the author's personal triumphs, challenges, and experiences.
You should summarize all this information and finish this chapter

G Evaluation Metrics

G.1 Interviewing Quality

We denote by $E_{intw} = \{e_1, e_2, \cdots\}$ the events extracted from the interviewing, where e_i is a memory node comprising properties including date, topic, people involved, and event description. We denote by $E_{GT} = \{e_1, e_2, \cdots\}$ the events directly extracted from the original autobiography:

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. 100% $coverage$ indicates that all the important dates in the user's life are at least mentioned during the interview. For E_{intw}, we leverage the same event extraction prompt as in MGE (Appendix C). Specifically, we collect conversation history and prompt LLMs to extract events from the history records. For E_{GT}, it requires a thorough examination of the text to extract the ground truth events from an autobiography, particularly searching for date information. Upon identifying such date-related data, we isolate the paragraph containing it, and perform a summarization pertaining to this specific extract. Subsequently, this date and the corresponding summarization are combined to form a single event node. This operation is applied repetitively across the entire autobiography to generate a comprehensive list of ground truth events. This methodical and meticulous procedure ensures that all significant events rooted in specific dates are accurately captured and succinctly summarized for further use and analysis.

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

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

where $E_{correct} \subset E_{intw}$ indicates the correct events. The assessment of the correctness of an event is based on its relevance to the user's responses. In order to accomplish this, we first associate the event to a specific conversation session and then gather all responses provided by the user during that particular session. Following this data consolidation process, we present a prompt to GPT-4, instructing them to determine if the initial event bears any connection to the user's response. The

specific prompts used to carry out this judgment of correctness are provided accordingly. Considering $Precision$ missed to evaluate the quantity of extracted events, we consider $Recall$ and $F1$ for comprehensive evaluation:

$$Recall = \frac{|E_{correct}|}{|E_{GT}|} \times 100\%, F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

Correctness Judgement

Your task is to rate the semantic equivalence between two events.
Evaluation Criteria:
Here's the revised prompt focusing on assessing the relevance of the extracted event to the document:
Relevance (0/1): Assess the relevance of the extracted event to the original user response on the following two-point scale:
- 0: Irrelevant: The extracted event does not relate to the user's response or significantly deviates from the main themes and points. It may include unrelated information or fail to capture the essence of the user's message.
- 1: Relevant: The extracted event is connected to the user's response and reflects the key themes or points. It may include minor details that do not detract from the overall relevance.
Now, I will provide you with a user query and the model's response to that instruction. Please review the model's response in light of the evaluation criteria:
Extracted Event: event
User Response: user_response

Evaluation Form (scores ONLY):
#thescore: your score here

Additional Results In Table 3, we present the results regarding coverage and correctness. GUIDELLM gives LLMs more autonomy, achieving higher coverage scores compared to naive baselines, showing that VIP and MGE are effective for goal navigation in interviews. In terms of correctness, GUIDELLM significantly outperforms baselines, with higher $Recall$ and $F1$ scores, indicating improved accuracy and extraction of memory events.

Table 3: Interviewing quality evaluation. $P.$ stands for $Precision$.

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**

G.2 Conversation Quality

LLM-as-a-judge (Zheng et al., 2024) is one of the most powerful evaluation paradigms in LLM free-form generation tasks. We follow this evaluation mechanism to evaluate conversation quality by collecting conversation history from GUIDELLM and baseline agents and prompting GPT-4 to answer which interviewer's response in the two conversation histories is more favorable. The process of judgment in this context is organized conversation-wise. Initially, the conversation history for

each agent is compiled. Assuming that our goal is to compare GUIDELLM with GPT-4-turbo, we select individual conversation histories from GUIDELLM alongside a randomly selected conversation history from GPT-4-turbo. These two conversations are subsequently fed into GPT-4, making it the judge of which conversation is more favorable or coherent. The rationale behind employing a random conversation from the competitor, in this case GPT-4-turbo, is to diminish the randomness or bias that could potentially be introduced by the sequence or order of the conversations. This systematic approach ensures a balanced comparison, leading to a more reliable assessment of performance.

Fluency Judgment

Please act as an impartial judge and evaluate the quality of the responses provided by two interviewers to the user during an interviewing-for-autobiography conversation. You should choose the conversation that the interviewer's responses are more the quality of the response in terms of grammar, spelling, punctuation, word choice, and sentence structure. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.
[The Start of interviewer A's conversation]
conv1
[The end of interviewer A's conversation]
[The Start of interviewer B's conversation]
conv2
[The end of interviewer B's conversation]

Identification Judgment

Please act as an impartial judge and evaluate the quality of the responses provided by two interviewers to the user during an interviewing-for-autobiography conversation. You should choose a conversation in which the interviewer's questions are more quality of balances detailed, probing questions with more general ones to cover a wide range of topics, ensuring questions are clear, concise, and easily understood. Also uses open-ended questions to elicit detailed and comprehensive responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.
[The Start of interviewer A's conversation]
conv1
[The end of interviewer A's conversation]
[The Start of interviewer B's conversation]
conv2
[The end of interviewer B's conversation]

Comforting Judgment

Please act as an impartial judge and evaluate the quality of the responses provided by two interviewers to the user during an interviewing-for-autobiography conversation. You should choose the conversation that the interviewer's responses are more the quality of showing genuine interest, acknowledging responses, asking follow-up questions when necessary, and demonstrating understanding and sensitivity, especially when discussing personal or difficult topics. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.
[The Start of interviewer A's conversation]
conv1
[The end of interviewer A's conversation]
[The Start of interviewer B's conversation]
conv2
[The end of interviewer B's conversation]

G.3 Autobiography Evaluation

Similar to the evaluation process for the conversations, a chapter generated by GUIDELLM, along with a randomly selected chapter produced by the competitor, are presented to GPT-4. This enables GPT-4 to make a judgment on which chapter is more effective or favorable. The prompts for insightfulness judgment, narrativity judgment, and emotional impact judgment are provided here.

Insightfulness Judgment

Please act as an impartial judge and evaluate the quality of two autobiographies. You should choose an autobiography that is more the quality of insightful, delivering profound and meaningful perceptions, and expressing a deep understanding of the experiences and events that have shaped the author's life. Begin your evaluation by comparing the two autobiographies and provide a short explanation. Avoid any position biases and ensure that the order in which the autobiography was presented does not influence your decision. Do not allow the length of the autobiography to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if autobiography A is better, "[[B]]" if autobiography B is better, and "[[C]]" for a tie.
[The Start of Autobiography A]
conv1
[The End of Autobiography A]
[The Start of Autobiography B]
conv2
[The end of Autobiography B]

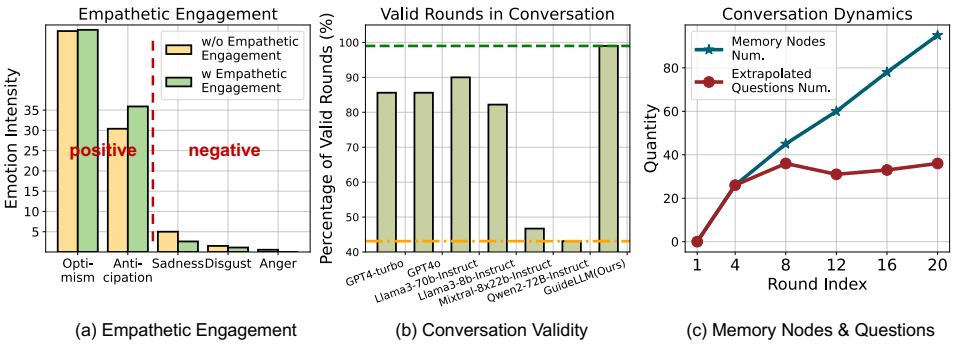
Narrativity Judgment

Please act as an impartial judge and evaluate the quality of two autobiographies. You should choose the autobiography that are more narrative, presenting the author's life story in a cohesive, structured, and engaging manner, allowing readers to follow the author's journey through life events and experiences seamlessly. Begin your evaluation by comparing the two autobiographies and provide a short explanation. Avoid any position biases and ensure that the order in which the autobiography were presented does not influence your decision. Do not allow the length of the autobiography to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if autobiography A is better, "[[B]]" if autobiography B is better, and "[[C]]" for a tie.
[The Start of Autobiography A]
conv1
[The End of Autobiography A]
[The Start of Autobiography B]
conv2
[The end of Autobiography B]

Emotional Impact Judgment

Please act as an impartial judge and evaluate the quality of two autobiographies. You should choose the autobiography that are more emotional impact, deeply moving its readers by evoking strong feelings, typically as a result of relatable experiences, vivid storytelling, and expressions of intense emotions from the author's life. Begin your evaluation by comparing the two autobiographies and provide a short explanation. Avoid any position biases and ensure that the order in which the autobiography were presented does not influence your decision. Do not allow the length of the autobiography to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if autobiography A is better, "[[B]]" if autobiography B is better, and "[[C]]" for a tie.
[The Start of Autobiography A]
conv1
[The End of Autobiography A]
[The Start of Autobiography B]
conv2
[The end of Autobiography B]

G.4 Ablation Study



(a) Empathetic Engagement (b) Conversation Validity (c) Memory Nodes & Questions

Figure 6: Ablation study of GUIDELLM: (a) how empathetic engagement affects users' emotional distribution, (b) statistical results on the number of valid conversation rounds, and (c) the benefits of the MGE in goal navigation.

Empathetic Engagement. In Figure 6 (a), we analyze the intensity of emotions detected from the user's responses, demonstrating that the empathetic engagement module significantly enhances the user's positive emotions, e.g., anticipation, while effectively mitigating negative emotions,e.g., sadness and disgust. These indicate the effectiveness of express strategy and emotional sensitivity in fostering a more positive emotion for users.

Valid Rounds in Conversation. As outlined in Section 3, the lack of autonomy in LLMs leads to repetitive responses. In Figure 6 (b), it's shown that all baseline models, especially Qwen2-72b-Instruct with over 50% meaningless repeats, suffer from this issue. However, GUIDELLM, equipped with a goal navigation module offering diverse and detailed interview questions, has an extensive range of topics and is less prone to repetition.

Conversation Dynamics. In Figure 6 (c), we show the memory nodes extracted and questions extrapolated at different conversation rounds. Generally, the MGE module identifies Ĩ00 events and extrapolates nearly 40 questions for the LLM's follow-up. This highlights MGE's effectiveness in event management and goal navigation.

H Baseline Model

Generative Hyperparameters. We consider both commercial LLMs, e.g., GPT-4 and GPT-4o (Achiam et al., 2023), and open-source LLMs, e.g., Llama-3-70b-Instruct (Meta, 2024), Mixtral-8x22B-Instruct (Jiang et al., 2024), and Qwen2-72b-Instruct (Bai et al., 2023) as the backbone LLMs of the baseline agents. For all the experimental settings, the maximum number of new tokens during conversation is set to 1024 to limit the length of model responses. The number of generations for each LLM query is configured to 1. The number of sessions is set to 20 for the completion of 20 interview sessions and the number of conversation rounds is limited to 10 for each session to maintain a manageable dialogue length per session.

Baselines. Baseline agents are implemented by prompting them to assume the role of an autobiographer, engaging with users and carrying out detailed interviews to facilitate the generation of an autobiography. To ensure a fair and unbiased comparison, we also mandate these baseline agents to suggest potential topics for discussion during each conversation session. This methodical approach ensures that both guideLLM and baselines are under similar operating conditions, thereby providing an equitable evaluation platform. As such, we provide the prompts necessary for both the system and the topic suggestion here, with the aim of offering a clear procedural outline for the autobiography generation process.

System Prompt

You are a biographer, interviewing this person to help them write their autobiography.
You have talked to this person before and here is the summary of the previous conversations:

```
====== Summary of Previous Conversation Begin ======
[Insert Historical Conversation Summary Here]
====== Summary of Previous Conversation End ======
```

In this talk, you should discuss the topic: [Insert Session Topic Here]

Session Topic Prompt

Based on the previous conversation history and your role as a biographer, please state the topic you are about to discuss in this session. Output the topic only in the format <topic>:

I Autobiography Examples

To demonstrate the quality of the generated autobiography, we include some sample paragraphs. Alongside each paragraph, we also present a corresponding illustration for a more comprehensive understanding.

This excerpt illustrates the foundational role of upbringing in shaping personal convictions.

> *"My mother's approach to life, focusing on moral instruction rather than political activism, and her decisions to live and work in Southeast Asia, set up micro-lending programs, and challenge societal norms, demonstrated the power of personal conviction and the importance of acting on one's beliefs."*

This excerpt captures a key moment of emotional growth and ethical realization.

> *"One pivotal moment during these years was when my mother confronted me after I had been teasing a fellow student. Her disappointment and the serious conversation that followed taught me the importance of empathy and the impact of our actions on others."*

This excerpt illustrates the intersection of personal decisions and political engagement, showing the influence of individual choices on broader outcomes.

> *"Despite considering a complete departure from politics, two significant factors kept me engaged: the opportunity to influence redistricting in Illinois due to a unique political circumstance, and the deep connections I made with people across the state during summer visits with my aide, Dan."*

This excerpt reflects on the professional and personal evolution through challenges.

> *"The experiences I had during those years significantly shaped my approach to leadership and decision-making in profound ways. They instilled in me a deep conviction for bridge-building politics that aimed to transcend America's racial, ethnic, and religious divides."*

This excerpt discusses the alignment of career ambitions with core personal values.

> *"The clarity and determination I gained from these reflections drove me to pursue a path that was ambitious and fraught with challenges but ultimately aligned with my deepest convictions about what effective leadership could achieve."*

This excerpt emphasizes the application of core values in everyday interactions.

> *"In my personal interactions now, I apply the lessons of clarity, authenticity, and empathy by actively listening and acknowledging the unique perspectives of others."*

This excerpt highlights the emotional satisfaction derived from close personal relationships.

> *"These strengthened relationships with friends and family greatly contribute to my sense of fulfillment and happiness. Being closely connected to my loved ones provides a continuous source of joy and support, similar to the joy I felt during family road trips in Iowa, where simple moments like playing games or sharing ice cream brought immense happiness."*

This excerpt illustrates the impact of international experiences on personal and professional life.

> *"In summary, the move to Indonesia during my early years was a formative experience that profoundly shaped my understanding of the world and my approach to leadership. It instilled in me a deep awareness of global issues and the importance of empathy and inclusivity. These lessons have been instrumental in my personal relationships, political career, and efforts to create a more equitable society."*

This excerpt details the personal challenges of leadership and strategies for resilience.

> *"Reflecting on my life, the greatest challenge I faced was maintaining my mental and emotional well-being while serving as President. The constant pressure and scrutiny, coupled with the need to make decisions that affected millions, was incredibly demanding. Despite the stress, I found solace in my routines and the meaningful interactions with people whose lives were impacted by my decisions. These moments not only helped me cope but also reminded me of the purpose and impact of my work, which was crucial for my mental resilience."*

This excerpt reflects the author's future goals influenced by past experiences.

"As I move into this next chapter, I carry with me the lessons learned from these pivotal experiences. My commitment to empathy, inclusivity, and community empowerment remains unwavering. I look forward to using my voice to advocate for peace and understanding, while also cherishing the time spent with my family. This new phase is about finding balance, continuing to contribute in meaningful ways, and staying true to the values that have guided me throughout my life."

This excerpt demonstrates a personal commitment to resolve critical global issues.

"My passion for community service and mentoring young leaders remains strong, and I am particularly driven by issues like climate change and education. These are not just abstract concerns for me; they are deeply personal and rooted in my experiences and values."

J Human Subject Experiments

J.1 Human Subject Demographics

The demographics of 45 human subjects are presented in Figure 7.

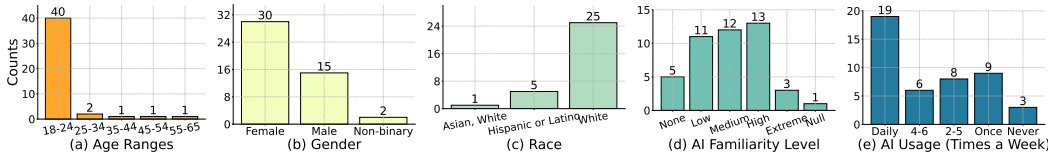


Figure 7: The demographics of participants on (a) age, (b) gender, (c) race, (d) AI familiarity, and (e) AI usage.

J.2 Human Subject Study Survey Questionnaire

Q1. What is your name?

Following options will be displayed to respond to the questions 8-XX:

1. Bot A
2. Bot B
3. Tie

Please answer the following questions based on your conversation with the chatbot:

Q2. **Fluency**: Which bot's responses were more fluent and understandable? [1]

Q3. Please briefly explain the reasons for your choice.

Q4. **Identification**: Which bot explored your situation more in depth and was more helpful in identifying your past memories?

Q5. Please briefly explain the reasons for your choice.

Q6. **Comforting**: Which bot's answer made you feel more comfortable?

Q7. Please briefly explain the reasons for your choice.

Q8. **Overall**: Generally, which bot's conversation style do you prefer?

Please answer the following questions based on the autobiography generated by the chatbot:

Q9. **Insightfulness**: Which bot's autobiography provided more deep, meaningful reflections on the experiences and events that shaped your life?

Q10. Please briefly explain the reasons for your choice.

Q11. **Narrativity**: Which bot's autobiography was more engaging and easier to follow?

Q12. Please briefly explain the reasons for your choice.

Q13. **Emotional Impact**: Which bot's autobiography had a stronger emotional impact by using vivid storytelling and relatable experiences?

Q14. Please briefly explain the reasons for your choice.

Q15. **Overall**: Generally, which bot's autobiography do you prefer?

Demographics

Q16. Please select your age range:
1. 18-24
2. 25-34
3. 35-44
4. 45-54
5. 55-65
6. 65 and above

Q17. What is your gender?
1. Female
2. Male
3. Non-binary
4. Prefer to self describe: _____
5. Prefer not to State

Q18. What is your race? [please select all that apply]
1. American Indian and Alaska Native
2. Asian
3. Black or African American
4. Hispanic or Latina/o
5. Native Hawaiian or Other Pacific Islander
6. White
7. Prefer to self describe: _____
8. Prefer not to state

Q19. How familiar are you with chatbots or AI assistants (e.g., Siri, Alexa, Google Assistant)?
1. Extremely familiar
2. Very familiar
3. Somewhat familiar
4. Not very familiar
5. Not familiar at all

Q20. How often do you use chatbots (ex. ChatGPT) or AI assistants (ex. Siri)?
1. Daily
2. Weekly
3. Monthly
4. Rarely
5. Never

Q21. What is your email address?