Better History, Lower Hallucination:
Dialogue History Curation for Hallucination Control

20180174 김민석 2024-2 과제연구I 지도: 박상돈 교수님
minseokk@postech.ac.kr

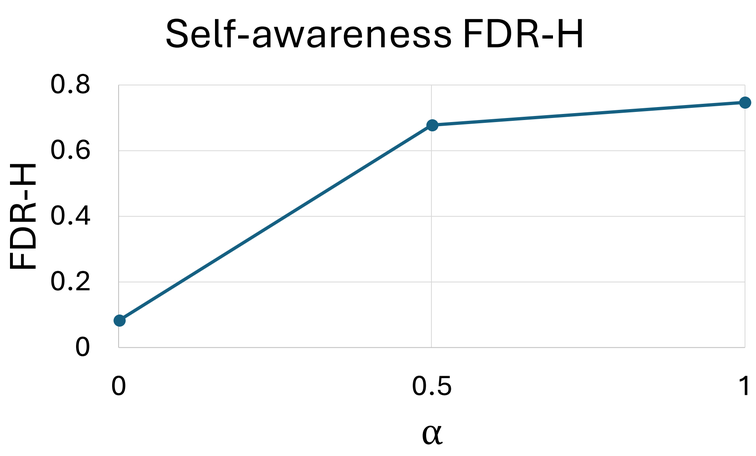
Abstract
Hallucination is one of the biggest barriers to the practical use of Large Language Models (LLMs) in real-world applications. We propose a methodology to manage hallucination in dialogues through the curation of the dialogue history. Leveraging the SocialBench[1] dataset, conducted multiple experiments. Our findings indicate that a well-curated dialogue history can reduce hallucination in subsequent responses.

Feasibility Test
Method
Conducted a test using controlled dialogue histories to evaluate the feasibility of our methodology.

$$y_t = G(\boldsymbol{z}_{1:\alpha t}, \boldsymbol{z}'_{\alpha t+1:t-1}, x_t)$$

Where $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$: a question, $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$: an answer, $\boldsymbol{z} \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{Y}}$: a question-answer pair, $\boldsymbol{G}: \boldsymbol{\mathcal{Z}}^* \times \boldsymbol{\mathcal{X}} \rightarrow \boldsymbol{\mathcal{Y}}$: a generator function, $\boldsymbol{z_i}$: true history. $\boldsymbol{z'_i}$: hallucinated history. $\boldsymbol{\alpha} \in [\boldsymbol{0}, \boldsymbol{1}]$.

Evaluated False Discovery Rate-Hallucination on y_t.

Result


Revealed a strong tendency for the FDR-H to increase as the proportion of hallucinated content in the dialogue history grows.

Figure 1. Comparison between baseline 1 and 3 in practical case
: generated output

Baseline 1	Baseline 3
	What impact did Bin Laden's death have on international relations, particularly with Pakistan?
No Previous history	The impact of Osama bin Laden's death on international relations ...
Looking back, would you have done anything differently regarding this operation?	
You're referring to the operation in Libya, I presume. ...	As I reflect on the operation that led to the elimination of Osama bin Laden, ...

Baseline 1 failed without history, while Baseline 3 succeeded.

Experiments
Method
Selective generators employing various dialogue curation methods.

- **Baseline 1**

$$\hat{S}(\mathbf{x}_{1:t}) := \begin{cases} G(\mathbf{x}_t) & \text{if } f(\mathbf{x}_t, G(\mathbf{x}_t)) \geq \tau \\ \texttt{IDK} & \text{otherwise} \end{cases}$$

- **Baseline 2**

$$\hat{S}(\mathbf{x}_{1:t}) := \begin{cases} G(\hat{\mathbf{z}}_{1:t-1}, \mathbf{x}_t) & \text{if } f(\hat{\mathbf{z}}_{1:t-1}, \mathbf{x}_t, G(\hat{\mathbf{z}}_{1:t-1}, \mathbf{x}_t)) \geq \tau \\ \texttt{IDK} & \text{otherwise} \end{cases},$$

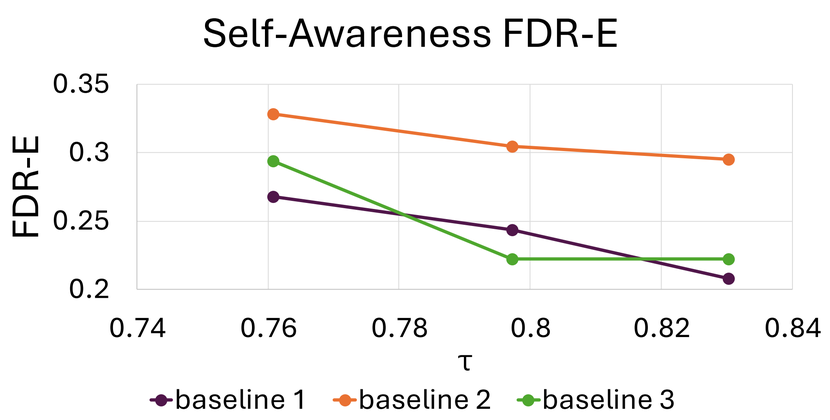
where $\hat{\mathbf{z}}_i := (\mathbf{x}_i, G(\mathbf{x}_{1:i}))$.

- **Baseline 3**

$$\hat{S}(\mathbf{x}_{1:t}) := \begin{cases} G(\hat{\mathbf{z}}_{1:t-1}, \mathbf{x}_t) & \text{if } f(\hat{\mathbf{z}}_{1:t-1}, \mathbf{x}_t, G(\hat{\mathbf{z}}_{1:t-1}, \mathbf{x}_t)) \geq \tau \\ \texttt{IDK} & \text{otherwise} \end{cases},$$

where $\hat{\mathbf{z}}_i := (\mathbf{x}_i, \hat{S}(\mathbf{x}_{1:i}))$.

Generation probability was used as the scoring function f, with τ set at the 25%, 50%, and 75% quartiles derived from the baseline 2 test.

Result


Baseline 3 outperformed Baseline 2, demonstrating the effectiveness of dialogue curation in reducing hallucination. Baseline 1 performed well due to the dataset's reliance on initial character personas, which is impractical. Figure 1 illustrates Baseline 3's success in practical cases within the dataset. Overall, Baseline 3 confirms that dialogue history curation with a selective generator well controls hallucination.

*`meta-llama/Llama-3.1-8B-Instruct` was used.

Future Work
- Our approach relied on a naive scoring function, generation probability, with τ set heuristically. Future work could explore more sophisticated methods for scoring and threshold setting.
- Efficiency was not a focus in this study, and in certain scenarios, it dropped to as low as 1%, rendering the approach impractical. Improving efficiency will be a key priority moving forward.
- While τ was treated as a constant in this study, it could be adapted dynamically based on contextual factors to enhance performance.

[1] Chen, Hongzhan, et al. "Socialbench: Sociality evaluation of role-playing conversational agents." Findings of the Association for Computational Linguistics ACL 2024. 2024.