Post-hoc Utterance Refining Method by Entity Mining for Faithful Knowledge Grounded Conversations

Yoonna Jang*, Suhyune Son*, Jeongwoo Lee*, Junyoung Son, Yuna Hur, Jungwoo Lim, Hyonseok Moon, Kisu Yang, and Heuiseok Lim†

Department of Computer Science and Engineering, Korea University
{morelychee, ssh5131, time79779, s0ny, yj727722, wjddn803, glee889, willow4, limhseok}@korea.ac.kr

Abstract

Despite the striking advances in recent language generation performance, model-generated responses have suffered from the chronic problem of hallucinations that are either untrue or unfaithful to a given source. Especially in the task of knowledge grounded conversation, the models are required to generate informative responses, but hallucinated utterances lead to miscommunication. In particular, entity-level hallucination that causes critical misinformation and undesirable conversation is one of the major concerns. To address this issue, we propose a post-hoc refinement method called REM. It aims to enhance the quality and faithfulness of hallucinated utterances by refining them based on the source knowledge. If the generated utterance has a low source-faithfulness score with the given knowledge, REM mines the key entities in the knowledge and implicitly uses them for refining the utterances. We verify that our method reduces entity hallucination in the utterance. Also, we show the adaptability and efficacy of REM with extensive experiments and generative results. Our code is available at https://github.com/YOONNAJANG/REM.

1 Introduction

The knowledge grounded conversation (KGC; also called knowledge grounded conversation, KGD), which is a subfield of the dialogue systems, is a task of generating human-like utterances by referring to specialized knowledge such as Wikipedia† (Zhao et al., 2022; Li et al., 2022c). The KGC task requires the ability to generate fluent and informative utterances based on source knowledge. Considering this ability, there has been numerous concurrent research on the KGC to build models that can have conversations with human-like expertise (Ma et al., 2020; Yu et al., 2022; Sluscher et al., 2022a; Liu et al., 2022; Sun et al., 2023; Xiao et al., 2023).

A colossal number of training corpora and parameters have brought the recent explosion of language generation performance in pre-trained language models (PLMs) (Brown et al., 2020; Raffel et al., 2020) and large language models (LLMs) (Thoppilan et al., 2022; Touvron et al., 2023; OpenAI, 2023). However, hallucination, which has been a chronic problem in natural language generation, is still cited as one of the biggest challenges remaining unsolved (Ji et al., 2023). In the KGC task, even though the ground truth knowledge is given, the models make the error of generating hallucinated utterances not faithful to the source knowledge (Li et al., 2022a; Ji et al., 2023).

Especially, entity-level hallucination, generating

Figure 1: An illustrated example of the utterances with entity-level hallucination (i.e., case 1 and 2) and desirable utterances in knowledge grounded conversation. In the source-faithful utterance, entities of the given source knowledge are included, making the utterance informative.

*These authors contributed equally to this work.
†Corresponding author.
†https://www.wikipedia.org/
names of entities that are incorrect or not present in
the source document (Nan et al., 2021), causes critical
misinformation and jeopardizes the flow of the
conversation (Das et al., 2023). As shown in Case 1
of Figure 1, the previous KGC models generate the
utterance containing information about the wrong
entity, which is not given in the knowledge. Further,
the generated utterance is excessively general while
not considering sufficient entities that align with
the context of the conversation, as shown in Case 2.
These deficiencies can undermine the helpfulness
of AI conversational models. Though previous re-
search has tried to mitigate them in general domain
discussion (Shuster et al., 2021), research to ad-
dress the entity-level hallucination in KGC remains
dark (Zhang et al., 2022).

To address these entity-level hallucination prob-
lems, we propose a post-hoc utterance refining
method by entity mining, called REM, for more
desirable and source-faithful KGC. REM can be
used to refine the unfaithful utterances generated
by previous models in a plug-and-play manner. To
refine utterances, we leverage the entity mining
method, which extracts the named entities to im-
plitically utilize key information in the knowledge
in a multi-tasking manner. With this simple but ef-
effective method, REM aims to mitigate entity-level
hallucination and lead to a more successful conver-
sation.

In order to measure the effectiveness of REM,
we conduct an extensive empirical evaluation.
First, to demonstrate the post-hoc refining ability
of REM, we experiment with refining the utter-
ances generated by baseline models on three KGC
datasets. We investigate the flexibility and valid-
ity of REM with cross-data experiments and ad-
versarial data refining experiments, respectively.
In addition, we conduct an ablation study and hu-
man evaluation to verify the effectiveness of our
method, showing the improvement of the source-
faithfulness score and entity coverage of refined
utterances. We also demonstrate the scalability
of REM by applying it to large language models.
From a case study comparing the refined results of
REM and baseline utterance, we demonstrate that
our REM model improves the source-faithfulness
in the utterances.

Our contributions are threefold: (1) We propose a
post-hoc refining method by implicitly mining key
entities in the knowledge for more source-faithful
conversation; (2) We show that our simple but ef-
fective method is adaptable to the existing models,
including large language models, in a plug-and-
play manner; (3) We substantiate that our method
reduces entity-level hallucination and accomplish
more desirable knowledge grounded conversation
with diverse experiments.

2 Related Work

2.1 Knowledge Grounded Conversation

In the task of knowledge grounded conversation
(KGC), the systems aim to generate an informati-
ve conversation based on specialized knowledge.
To support research in this area, publicly avail-
able datasets for the KGC task have been devel-
oped (Gopalakrishnan et al., 2019; Moon et al.,
2019; Shuster et al., 2022b). These datasets fo-
cus particularly on generating informative con-
versations on specific topics (Dinan et al., 2018;
Zhou et al., 2018; Jang et al., 2022). Building
upon these KGC datasets, there has been active
research to generate contextually consistent utter-
ances while utilizing the source knowledge. Kim
et al. (2020); Adolphs et al. (2021); Wang et al.
(2021); Zhang et al. (2023); Feng et al. (2023) in-
corporate knowledge and context selectors to filter
out irrelevant knowledge sentences and redundant
dialogue history, respectively. Additionally, (Niu
et al., 2023) propose the history-adapted knowledge
copy (HAKC) network, which selectively chooses
context-aware knowledge to maintain dialogue co-
herence. Likewise, diverse research is widely con-
ducted in KGC tasks.

2.2 Knowledge Hallucination in KGC

Despite the remarkable advancements, KGC sys-
tems are known to suffer from knowledge hal-
locination generating unfaithful utterances to
source knowledge (Dziri et al., 2021b). To ad-
dress hallucination, Shuster et al. (2021) pro-
pose neural-retrieval-in-the-loop architectures to
improve knowledge-ability consisting of retrievers,
rankers, and encoder-decoders. Additionally, Li
et al. (2022b) proposes a method that utilizes entity
and relation information from a knowledge graph
to generate more faithful utterances. However, in
KGC, there is a lack of studies on entity-level hal-
lucination, which directly identifies and controls
the generation of unfaithful utterances (Das et al.,
2023). Although Dziri et al. (2021a) explore the
study on entity-level hallucination, this study only
considers cases where the source of knowledge is
a knowledge graph. Therefore, there is a need for research on entity-level hallucination in KGC tasks where the source knowledge is provided.

2.3 Utterance Refining Methods

Refining methods have been studied to improve the generation results when they are not satisfactory or not in the desired intention. (He, 2021; Geng et al., 2021; Sun et al., 2022; Bao and Zhang, 2023) Especially in dialogue systems, the researches aim to improve the quality of response by applying the refining method have been proposed (Tran and Nguyen, 2018; Posokhov et al., 2022; Wang et al., 2022). To tackle the limitation of generating uninformative or not engaging utterances (Shen et al., 2018), Weston et al. (2018) retrieves the utterance and refines it by regenerating the retrieved utterance. Moreover, for fascinating dialogue systems, the ability to generate personalized responses according to the user’s dialogue history is required (Cao et al., 2022). To address this problem, Zhong et al. (2022) refines the user dialogue history to extract valuable information. Furthermore, they generate personalized dialogue by utilizing refined history information. Also, Song et al. (2020) adopts a rewriting method for persona-consistent dialogue generation. In this study, our goal is to reduce entity-level hallucination, which is critical in informative conversation (Corbelle et al., 2022), with the refining method.

3 Proposed Method

Our proposed REM aims to refine the model-generated utterance to be more faithful to the source knowledge, as shown in Figure 3. The utterance is filtered by the faithfulness scoring function, whether to be refined or not. We train the pre-trained language model with an encoder-decoder structure in the REM method. In the REM model, the entity miner extracts the named entities from the source knowledge and learns them implicitly. The utterance generator makes more faithful utterances based on the key information extracted by the entity miner. In this section, we formulate a KGC task and our proposed utterance refining method, and describe the REM model and its training process.

3.1 Knowledge Grounded Conversation

As depicted in Figure 2 (a), the KGC models generate the informative utterance \( \hat{u} \) considering both the conversational context (or dialogue history) \( c \) and the corresponding source knowledge \( k \) as follows:

\[
p_\theta(\hat{u}|k, c) = \prod_{t=1}^{n} p_\theta(\hat{u}_t|\hat{u}_{<t}, k, c),
\]

where \( n \) indicates the max sequence length of the utterance, and \( \theta \) denotes the KGC model parameter.

3.2 Post-hoc Utterance Refining

While KGC models aim to generate informative utterances, model-generated utterance \( \hat{u} \) may not reflect the input source knowledge faithfully, as shown in Figure 1. In this context, we first filter out the utterance which is not faithful to the source. Then we renovate the given utterance into the refined utterance \( \tilde{u} \) which represents the intended knowledge better by utilizing the model-generated utterance \( \hat{u} \) and its corresponding source knowledge \( k \) from KGC models as follows:

\[
p_\psi(\tilde{u}|k, \hat{u}) = \prod_{t=1}^{n} p_\psi(\tilde{u}_t|\tilde{u}_{<t}, k, \hat{u}),
\]

where \( \psi \) denotes the REM model parameter.

Faithfulness Filtering To filter the model-generated utterance \( \hat{u} \), we quantify the source-faithfulness score, which indicates how \( \hat{u} \) is consistent with the source knowledge \( k \). To this end, we adopt the DAE (Goyal and Durrett, 2020) as
Figure 3: The architecture of REM model, which consists of entity miner and utterance generator. The utterance $\tilde{u}$ with a lower source-faithfulness score than threshold $\tau$ enters as input to REM with source knowledge $k$. The entity miner extracts entities $e$ from $k$, in white circles, while the utterance generator refines the $\tilde{u}$ with the extracted entity $e$.

a scoring function to estimate the entailment between the source knowledge and $\tilde{u}$ considering the dependency arc-level consistency of the utterance.

For efficiency, we only refine the utterances that have lower scores than the threshold according to the source-faithfulness score, as follows:

$$\tilde{u} = \begin{cases} 
\text{REM}(k, \tilde{u}) & \text{if } \text{score}(k, \tilde{u}) < \tau \\
\tilde{u} & \text{otherwise}, \end{cases} \quad (3)$$

where $\text{score}(\cdot)$ denotes the scoring function (i.e., DAE) and $\tau$ is a threshold value between 0 and 1. When $\tau$ is 1, all utterances are re-generated as $\tilde{u}$; otherwise, when $\tau$ is 0, all utterances are not refined. We utilize this filtering module only in the inference step, not in the training step.

### Refining Utterance by Entity Mining (REM)

Our refining method, REM, is decomposed into two sub-modules that are (i) entity miner and (ii) utterance generator, as shown in Figure 2. The refining task is formulated as follows:

$$p_\phi(\tilde{u} | k, \tilde{u}) \propto p_\psi(e | k, \tilde{u}) p_\psi(\tilde{u} | e, k, \tilde{u})$$

(4)  

These two modules are trained in a multi-tasking manner with parameter $\psi$. The former mines entities $e$ from source knowledge and learns entity-level knowledge implicitly. The latter learns to generate the utterance $\tilde{u}$ with the implicitly mined entities, which is key information of source knowledge, reducing entity-level hallucinations and producing more source-faithful utterances.

### 3.3 Training Objectives

The training objective of REM mainly falls into named entity recognition (NER) by entity miner and utterance re-generation (GEN) by utterance generator.

To predict the named entities, the NER head is attached on top of the encoder and mines the entities inside the knowledge $k$. It makes the model learn the essential entity information, enhancing the interpretation of the source knowledge. To train NER, we tag named entities with one of the following entity types {‘LOC’, ‘PER’, ‘ORG’, ‘MISC’} using the off-the-shelf entity tagging module\(^2\) for the source knowledge in the data.

The NER loss $\mathcal{L}_{NER}$ is only defined for the NER label $l_i$ tagged for each token $k_i$ in $k$ and minimized during training as follows:

$$\mathcal{L}_{NER} = -\frac{1}{N_k} \sum_{i=1}^{N_k} l_i \log p(k_i | \tilde{u}), \quad (5)$$

where $N_k$ denotes the token length of the source knowledge.

The language modeling head (GEN) learns to refine the given model-generated utterance $\tilde{u}$ along with $k$ in an auto-regressive manner while extracting entities $e$. The model learns to compose the information of the inputs and extracted entities from

\(^2\)https://github.com/flairNLP/flair
the encoder hidden states. It is trained by minimizing the following loss function:

$$\mathcal{L}_{GEN} = -\frac{1}{N_u} \sum_{i=1}^{N_u} \log p(u_i | u_{<i}, k, \hat{u}; e), \quad (6)$$

where $e$ is implicitly mined in the encoder during training.

The final training objective for a REM trained on both tasks is the following, where $\lambda_n$ is the training hyperparameter:

$$\mathcal{L}_{REM} = \lambda_1 \mathcal{L}_{NER} + \lambda_2 \mathcal{L}_{GEN} \quad (7)$$

### 4 Experimental Settings

In this section, we introduce the dataset, the automatic evaluation metric, and the baseline models used in the experiments. The implementation details are attached in Appendix C.

#### 4.1 Datasets

We utilize three datasets as our testbed, but we only use instances for training and testing where ground truth knowledge is given in the data.

FoCus (Jang et al., 2022), which considers both knowledge and persona in conversation, has been released. In FoCus, machine generates utterances with customized and knowledgeable utterances about world landmarks. Wizard of Wikipedia (WoW) (Dinan et al., 2018), which is a widely used benchmark, consists of Wizard and Apprentice talking to each other on various topics in Wikipedia. Another dataset, CMUDoG (Zhou et al., 2018) includes conversations between two speakers discussing different aspects of a specific movie, such as information, plot, etc. The data statistics are presented in Appendix B.

#### 4.2 Automated Evaluation Metrics

We evaluate the refining performance of REM with automated metrics in three criteria: source-faithfulness, reference matching, and diversity. To evaluate source-faithfulness, we adopt DAE, entity coverage (EC), and entity type coverage (TC).

DAE (Goyal and Durrett, 2020) is used to assess whether the system accurately reflects the facts from the given knowledge when generating utterances at the level of dependency arcs. Furthermore, we utilize two entity-level metrics (EC and TC), following subsequent paragraphs, to evaluate the source-faithfulness in terms of the entity. For the reference matching n-gram score, we utilize chrF (Popović, 2015), ROUGE-L (Lin, 2004) and SacreBLEU (Post, 2018) to evaluate how close the generated hypothesis is to the ground-truth answer in the test set. For diversity evaluation, Distinct-N$^3$ (Li et al., 2015) is adopted.

**Entity Coverage and Entity Type Coverage** We compute the entity coverage (EC) and entity type coverage (TC) to quantify the extent of correct entities and context coherence, respectively. To compute EC and TC, we extract the named entities from the generated (model-predicted) utterance $u_p$ (including $\hat{u}$ and $\hat{u}$), source knowledge $k$, and ground truth utterance $u$, respectively, using the off-the-shelf NER model (Schweter and Akbik, 2020) (The distribution of named entity types are described in Appendix B). First, we evaluate the ratio of the named entities residing in the generated utterance $u_p$ by comparing it with the ground truth utterance $u$ as follows:

$$EC = \frac{\mathcal{N}(E_k \cap E_u \cap E_{u_p})}{\mathcal{N}(E_k \cap E_u)}, \quad (8)$$

where $\mathcal{N}(l)$ is the number of values in list $l$ and $E_k$ is the set of named entities in source knowledge. $E_u$ and $E_{u_p}$ indicate the set of named entities in ground truth utterance and generated utterance, respectively.

In addition, we evaluate TC, as existing models produce fluent text but demonstrate low context coherence. To identify the intent of context and the information that should be included in the response, we leverage the named entity types in the ground truth utterance $u$. For example, if the entity type ‘LOC’ has the highest proportion among the named entities in $u$, it indicates that the conversation is focused on location-related information. In such cases, the model should generate responses that are relevant to the location to ensure coherence and relevance in the conversation.

To this end, we compute the ratio of name entity types extracted from the generated utterance $u_p$ compared to the ground truth utterance $u$ in each named entity type.

$$TC = \frac{1}{|T|} \sum_{t} \frac{\mathcal{N}(E_{u_p}^t)}{\mathcal{N}(E_u^t)}, \quad (9)$$

where $T$ is the set of named entity types \{‘LOC’, ‘PER’, ‘ORG’, ‘MISC’\}.3

---

4.3 KGC Baselines

BART BART (Lewis et al., 2020a) is Transformer-based (Vaswani et al., 2017) pre-trained model with the encoder-decoder structure. It utilizes various noising methods and in-filling schemes during pre-training. We fine-tune BART on three KGC datasets and use its predicted utterance for train, validation, and test set.

INFO INFO (Lim et al., 2022) leverages RAG (Lewis et al., 2020b) to enhance the factuality of the generation results in FoCus dataset. It retrieves a large number of knowledge documents and generates informative conversations based on them.

EDMem We utilize and refine the utterances generated by EDMem (Zhang et al., 2022). EDMem is pre-trained on Wikipedia documents to learn entity embeddings and incorporates entity knowledge for entity-intensive dialogue generation.

ITDD ITDD (Li et al., 2019) is the best-performing method on the CMUDoG dataset. With a two-pass decoder inspired by the human cognitive process, it improves context coherence and knowledge correctness.

5 Results and Analysis

In this section, we report the experimental results and analysis. We first discuss the post-hoc refining ability of our method in §5.1. Then, we evaluate the flexibility of REM through cross-data refining experiments in §5.2. We also conduct experiments on refining adversarial utterances in §5.3, and perform an ablation study in §5.4. Furthermore, we provide the results of human evaluation in §5.5, and present a case study in §5.7. Finally, we analyze the application of REM to large language models in §5.6.

5.1 Post-hoc Refining Ability

To demonstrate the post-hoc refining ability, we evaluate the refining performance of REM with vanilla PLM fine-tuned on three datasets and existing models from previous studies. The results are in Table 1 where REMbase and REMlarge refer to the models trained using the REM method on the BART-base and BART-large (Lewis et al., 2020a).

Fine-tuned Baselines The comparison results of REM with the vanilla PLM fine-tuned on three datasets are presented at the top of Table 1. We fine-tune the BART-base model with each dataset and refer to it as BARTbase. To the generated outputs of BARTbase, we evaluate the refining performance with REMbase. The results demonstrate that the REM leads to improvements not only in the scores of source-faithfulness metrics but also in the reference matching metrics. We also show the ability of REM that refines the utterances of the larger model in Appendix F.

In contrast, REM shows a tendency to decrease the performance of diversity metrics. This indicates that in BARTbase-generated utterances, there are tokens that contribute to increased diversity but also lead to hallucination. These hallucinated tokens are subsequently refined through the REM process. In particular, the significant improvements in EC and TC metrics suggest that entities that cause entity-level hallucination are refined during the refining process with REM.

Existing Baselines To investigate the adaptability of REM, we compare the refining performance with previous studies, and the results are annotated with † symbol in Table 1. We re-implement EDMem, ITDD, and INFO for fair comparison with REMlarge. The results reveal that REMlarge exhibits an improvement in source-faithfulness performance. Additionally, it demonstrates enhanced generation performance in the reference matching score, excluding INFO. The reason is that INFO utilizes a large number of knowledge documents for generation, whereas REM generates utterances based on the given ground truth knowledge within the dataset. Furthermore, while the source-faithfulness score increases, the diversity score decreases. This result discusses that REM removes tokens that enhance diversity but also contribute to entity hallucination, similar to the results observed with vanilla PLM.

5.2 Cross-data Utterance Refining

To inspect the flexibility of REM, we conduct cross-data experiments, and the results are shown in Table 2. REM trained on CMUDoG consistently underperforms in source-faithfulness and reference matching scores on the other two datasets. When refining FoCus test set with REM trained on CMUDoG, the performance significantly decreases in all metrics. This tendency is similarly shown in the results of WoW test set. Likewise, the REM trained
Table 1: Experimental results of baseline utterance refining with REM. The top half shows the results of refining (1) the fine-tuned baselines, and the bottom half shows the results of refining (2) the existing baselines. Numbers in boldface indicate higher scores. † denotes the evaluation settings of the existing models, and the vanilla setting (without †) denotes the evaluation settings of our fine-tuned baseline. The utterances are filtered with $\tau = 0.5$.

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Source-faithfulness</th>
<th>Reference Matching</th>
<th>Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EC (%)</td>
<td>TC (%)</td>
<td>DAE</td>
</tr>
<tr>
<td>FoCus</td>
<td>BART$_{base}$</td>
<td>7.54</td>
<td>13.85</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>+ REM$_{base}$ (Ours)</td>
<td>23.40</td>
<td>20.32</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>BART$_{base}$</td>
<td>14.23</td>
<td>8.80</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>+ REM$_{base}$ (Ours)</td>
<td>14.57</td>
<td>9.03</td>
<td>0.87</td>
</tr>
<tr>
<td>WoW</td>
<td>BART$_{base}$</td>
<td>3.40</td>
<td>3.90</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>+ REM$_{base}$ (Ours)</td>
<td>4.95</td>
<td>7.55</td>
<td>0.40</td>
</tr>
<tr>
<td>CMUDoG</td>
<td>BART$_{base}$</td>
<td>14.23</td>
<td>8.80</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>+ REM$_{base}$ (Ours)</td>
<td>14.57</td>
<td>9.03</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>INFO</td>
<td>0.05</td>
<td>24.40</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>+ REM$_{large}$ (Ours)</td>
<td>0.11</td>
<td>25.48</td>
<td>0.86</td>
</tr>
<tr>
<td>WoW†</td>
<td>EDMem</td>
<td>0.00</td>
<td>3.50</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>+ REM$_{large}$ (Ours)</td>
<td>4.51</td>
<td>6.51</td>
<td>0.77</td>
</tr>
<tr>
<td>CMUDoG†</td>
<td>ITDD</td>
<td>0.00</td>
<td>1.42</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>+ REM$_{large}$ (Ours)</td>
<td>1.01</td>
<td>3.96</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 2: The result of cross-data experiments with REM$_{large}$. All utterances are refined with $\tau = 1.0$.

<table>
<thead>
<tr>
<th>Data</th>
<th>Source-faithfulness</th>
<th>Test Utterance</th>
<th>Source-faithfulness</th>
<th>Test Utterance</th>
<th>Reference Matching</th>
<th>Source-faithfulness</th>
<th>Test Utterance</th>
<th>Reference Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>FoCus</td>
<td>50.64</td>
<td>29.88</td>
<td>0.84</td>
<td>50.47</td>
<td>46.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WoW</td>
<td>0.03</td>
<td>8.35</td>
<td>0.03</td>
<td>9.24</td>
<td>8.65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMUDoG</td>
<td>17.02</td>
<td>16.20</td>
<td>0.29</td>
<td>21.32</td>
<td>15.72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WoW†</td>
<td>48.04</td>
<td>34.19</td>
<td>0.87</td>
<td>41.91</td>
<td>36.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMUDoG†</td>
<td>5.94</td>
<td>5.82</td>
<td>0.32</td>
<td>16.53</td>
<td>16.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WoW†</td>
<td>18.93</td>
<td>12.71</td>
<td>0.75</td>
<td>33.03</td>
<td>29.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMUDoG†</td>
<td>17.58</td>
<td>12.38</td>
<td>0.75</td>
<td>34.09</td>
<td>28.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WoW†</td>
<td>3.33</td>
<td>4.98</td>
<td>0.23</td>
<td>12.07</td>
<td>12.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMUDoG†</td>
<td>3.21</td>
<td>10.08</td>
<td>0.44</td>
<td>15.37</td>
<td>10.92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WoW†</td>
<td>5.99</td>
<td>9.24</td>
<td>0.42</td>
<td>15.55</td>
<td>11.43</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Experimental results of adversarial data refining. ‘ADV’ denotes the adversarially edited utterances by ChatGPT. All utterances are refined with $\tau = 1.0$.

5.3 Adversarial Data Refining

To investigate the validity of REM, we have prompted ChatGPT (OpenAI-Blog, 2022) to generate the synthetic data by adversarially changing the ground truth utterance$^4$. We demonstrate the performance of REM by refining the utterances that are distorted to be unfaithful to the source knowledge; all data are refined with ($\tau = 1.0$). As presented in Table 3, compared to the adversarially synthesized

$^4$The prompts are shown in Appendix E.
data (ADV), both source-faithfulness and reference matching scores increased significantly after refining. ROUGE-L score of the ADV is slightly higher in CMUDoG, which is likely due to the characteristic that the utterances of CMUDoG have much shorter lengths than other datasets (Dziri et al., 2022a).

5.4 Ablation Study
To explore the effect of entity miner, we conduct an ablation study of our proposed method. In Table 4, we present the results on the FoCus dataset, which has the most informative conversations in §5.2. We compare the REM model trained with multi-task learning and only trained with utterance refining (w/o NER).

In case the model is trained without entity miner, the results show a performance decrease across the board except in one case. Our proposed REM method, which learns essential information with entity miner, can be demonstrated to be effective for source-faithfulness and reference matching scores. The DAE score of the large models shows better performance without entity miner. The large model shows unstable results in §5.3. However, a significant improvement in the performance of the REM model is shown, with or without entity miner, compared to the baseline utterance. This suggests that refining helps increase the quality and source-faithfulness of fatal utterances.

5.5 Human Evaluation
To qualitatively evaluate the results before and after refining, we perform a human evaluation on 100 randomly sampled utterance examples, each from three datasets. One example contains an utterance produced by the baseline and the refined utterance by REM, and we evaluate them with three criteria: 1) fluency, 2) source-faithfulness, and 3) paraphrasing. The first criterion assesses naturalness by measuring the fluency of the generated results. The second criterion measures source-faithfulness by assessing whether the sentence is factually consistent with the given knowledge. In addition, the third criterion evaluates whether the given knowledge has been appropriately reorganized and incorporated into the utterance rather than just copied and pasted. We provide the questionnaire used in Appendix G.

The results of the comparative human evaluation of baseline utterances with scores below the threshold and utterances after refining are shown in Figure 4 ($\tau = 0.5$). Across all data and criteria, the refined utterances win in most cases. Especially in FoCus, refined utterances do not lose to the utterance before refining. REM performs the worst in CMUDoG, which is similar to the experiment in Section 5.1. The dissimilarity between the CMUDoG and the other two datasets can be attributed to its predominant resemblance to a general utterance instead of the knowledge-grounded conversations. Nevertheless, we qualitatively demonstrate the effectiveness of REM with a significantly higher number of winning cases.

5.6 REM-LLM
To show the scalability of our method to large language models (LLMs), we apply REM method to the prompt of not-tunable LLM (i.e., ChatGPT) beyond fine-tuning the models. In this experiment, large language models with (LLM$_{REM}$) and without (LLM) the REM method refine the utterance of the KGC baseline model (BART$_{base}$) and ChatGPT (LLM$_{KGC}$). LLM is required to refine the utterance considering the given source knowledge, while LLM$_{REM}$ is made to modify the utterance by extracting key entities from the knowledge and utilizing them. We attach the prompts used in Ap-
### 5.7 Case Study

To analyze the generative results of REM, we present the examples of three datasets in Table 6. In FoCus example, the utterance of baseline answers only provides the information of year, but with the imperfect entity mention ‘Per’. However, after being refined with REM, the utterance provides more specific knowledge with correct entity mentions, even utilizing the persona knowledge. In WoW example, the baseline utterance provides knowledge that is not given in the knowledge, but REM gives source-faithful information. In CMUDoG example, REM provides more detailed factual knowledge in the utterance than in the baseline utterance. We show more examples also in Appendix I.

Table 5: The result of REM with the large language model (denoted with +) refining the KGC utterances of BART_base and large language model (LLM_KGC). The utterances are filtered with $\tau = 0.5$.

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Source-faithfulness</th>
<th>Ref. Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data Model</td>
<td>EC</td>
<td>TC</td>
</tr>
<tr>
<td>FoCus</td>
<td>BART_base</td>
<td>7.56</td>
<td>13.85</td>
</tr>
<tr>
<td></td>
<td>+ LLM</td>
<td>23.54</td>
<td>20.14</td>
</tr>
<tr>
<td></td>
<td>+ LLM_REM</td>
<td>25.28</td>
<td>22.18</td>
</tr>
<tr>
<td>WoW</td>
<td>BART_base</td>
<td>14.23</td>
<td>8.79</td>
</tr>
<tr>
<td></td>
<td>+ LLM</td>
<td>14.34</td>
<td>8.89</td>
</tr>
<tr>
<td></td>
<td>+ LLM_REM</td>
<td>14.57</td>
<td>9.12</td>
</tr>
<tr>
<td>CMUDoG</td>
<td>LLM_KGC</td>
<td>3.40</td>
<td>3.90</td>
</tr>
<tr>
<td></td>
<td>+ LLM</td>
<td>4.93</td>
<td>6.07</td>
</tr>
<tr>
<td></td>
<td>+ LLM_REM</td>
<td>7.36</td>
<td>11.39</td>
</tr>
</tbody>
</table>

As shown in Table 5, both LLM and LLM_REM show improved source-faithfulness performance, especially in CMUDoG and FoCus. LLM_REM is able to achieve the most significant improvement by entity mining and refining the utterance with key entities. This demonstrates that when prompting with LLMs, more source-faithful utterances can be produced with REM by asking LLMs to find the important parts and use them for making responses.

6 Conclusion

In this work, we proposed REM, a post-hoc refining method for improving the source-faithfulness of the utterances in the knowledge grounded conversation (KGC). REM enabled simple but effective refining by extracting entities from the source knowledge given in the KGC task. It makes the model learn the key information implicitly and use it for refining the utterance more faithful to the source knowledge. We presented extensive experiments applying REM in a plug-in-play method to various model-generated outputs, showing increased source-faithfulness and entity coverage after refining. Also, qualitative analysis and human evaluation proved the refining efficacy of REM. We verified that our proposed method could be utilized with the prompt of the large language models.

---

Table 6: Examples of refined utterances by REM on three datasets.

<table>
<thead>
<tr>
<th>[Knowledge]</th>
<th>Designed by French architect Paul Andreu, the NCPA opened in 2007 and is the largest theatre complex in Asia.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Persona Knowledge]</td>
<td>I love art. I’m interested in architecture. I love lakes. I want to go on a trip. I want to visit China.</td>
</tr>
<tr>
<td></td>
<td>+ REM_base: National Center for the Performing Arts (NCPA) opened in 2007 and is the largest theatre complex in Asia. If you love art, you have to come see it with your own eyes!</td>
</tr>
</tbody>
</table>

---

Table 6: Examples of refined utterances by REM on three datasets.

<table>
<thead>
<tr>
<th>[Knowledge]</th>
<th>Among other honors, he has won nine Goyas, two European Film Awards and an Oscar.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Utterance]</td>
<td>BART_base: I’ve seen it a few times. It was directed by Alejandro Amená</td>
</tr>
<tr>
<td></td>
<td>+ REM_base: of course! He has won nine Goyas, two European Film Awards, and an Oscar! That’s a lot of awards for a director!</td>
</tr>
</tbody>
</table>

---

Table 6: Examples of refined utterances by REM on three datasets.

<table>
<thead>
<tr>
<th>[Knowledge]</th>
<th>The Asgardian Loki encounters the Other, the leader of an extraterrestrial race known as the Chitauri. ..., Loki takes the Tesseract and uses his scepter to enslave Selvig and a couple of other agents, including Clint Barton, to aid him in his getaway.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Utterance]</td>
<td>BART_base: you should watch the avengers 2 and 3</td>
</tr>
<tr>
<td></td>
<td>+ REM_base: you should watch the avengers 2 and 3, they are really good as well. what did you think of the scene where Loki takes the Tesseract?</td>
</tr>
</tbody>
</table>

---

Table 6: Examples of refined utterances by REM on three datasets.

---

Table 6: Examples of refined utterances by REM on three datasets.
Limitations

Our research proposed a method that aims to refine the non-source-faithful utterances in the knowledge grounded conversation. Though we address the problem of models not reflecting even the ground truth knowledge given, it is difficult to solve if the retriever does not give accurate knowledge from the in-the-wild setting. Therefore, the performance of the refiner may depend on the retriever’s performance. As it is beyond the scope of this paper, we leave it as future work. When filtering the utterances by source-faithfulness score, we only utilized DAE, but other scores or classifiers (Dou et al., 2022; Manakul et al., 2023) can be adopted for filtering according to the data or domain. There is an area of research in detecting critical errors in generated results in neural machine translation. Similarly, in the dialogue tasks, the better the classifier performs in detecting hallucinated utterances, the better the performance of the refiner will be. In addition, we expect that existing baseline models could be further improved if trained with REM in an end-to-end manner, so we leave it as future work.

Ethics Statement

The datasets used in our work are from previously published papers, so we do not attach privacy or ethical issues to the dataset. At inference, we will make the model not generate tokens in a list of several stopwords utilizing Hatebase, Silva et al. (2016), etc., to avoid generating harmful utterances that the model may have learned during pre-training. As we are aware that excessive computational energy used to train models stimulates environmental problems, we adopt the multi-task learning method for training entity miner and utterance generator instead of having a separate entity miner model to improve efficiency. Also, we distribute all codes and model checkpoints so they do not have to be trained again. We believe that our refining method can contribute to mitigating the entity-level hallucination of model-generated utterances and preventing the misuse of AI systems.

Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2018-0-01405) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-00-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). Also, this work was supported by NCsoft NLP Center.

References


A Dataset Details

FoCus (Jang et al., 2022) is a dataset where a human and a machine take turns having a conversation about a specific landmark. This dataset consists of 14,452 conversations, with 173,424 utterances. We use the validation set, which has 1,445 conversations and 17,340 utterances, for the experiment, as the official test set does not provide ground truth knowledge and utterances.

Wizard of Wikipedia (WoW) (Dinan et al., 2018) is a conversational dataset based on Wikipedia articles on various topics. The dataset covers 1,365 topics and consists of 22,311 conversations with a total of 201,999 utterances, which are divided into 166,787 for train, 17,715 for validation, and 17,497 for test. We used a random split version of the validation and test set. Only for evaluating ED-Mem (Zhang et al., 2022), we follow its test setting, KILT (Petroni et al., 2021), a variant of the WoW dataset.

CMUDoG (Zhou et al., 2018) is a knowledge grounded conversation dataset with two speakers conversing based on movie Wikipedia articles. Unlike the WoW and FoCus, where only one of the two speakers has access to the knowledge, both speakers of this dataset have access to the content of the article. It has a resemblance to a generic dialogue compared to the other two datasets. It is composed of a total of 4,112 conversations with an average of 21.43 turns.

B Distribution of Named Entity-Type

We show the entity type distribution of all datasets, the train, development, and test sets used for fine-tuning the baseline model BART base, in Figure 5. WoW has the evenest distribution, and FoCus has the second most even distribution. CMUDoG has the most “PER” entity class among the other three classes.

C Implementation Details

When implementing REM, we adopt BART (Lewis et al., 2020a), which has an encoder-decoder structure, and train the entity mining and utterance generation tasks. We experiment with BART base of 140M parameters and BART large of 406M parameters. We train REM on the pairs of model-predicted utterances, generated by the fine-tuned BART baseline, and reference utterance. We implement the models by exploiting Pytorch (Paszke et al., 2019) and HuggingFace (Wolf et al., 2019) with a fixed seed for reimplementation. The models are trained with a learning rate of 6.25e-5 for 10 epochs with early stopping. AdamW (Loshchilov and Hutter, 2017) is used as the optimizer. We set the train batch size to 8 with a gradient accumulation of 32. The time required for training is about 2 hours per epoch on a single RTX-6000 GPU. For decoding, we used a beam size of 5, min length of 32, max length of 512, top k of 50, and no-repeat n-gram size of 2. NER loss weight $\lambda_1$ and GEN loss weights $\lambda_2$ are set as 0.5:1, 0.3:1, and 0.7:1 for WoW, CMUDoG, and FoCus, respectively, according to the preliminary study. We will open the source codes which we used for the experiments after the review process.

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>ACC (NER)</th>
<th>FI (ACC)</th>
<th>PPL (GEN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FoCus</td>
<td>+ REM base</td>
<td>0.991</td>
<td>0.944</td>
<td>3.778</td>
</tr>
<tr>
<td>FoCus</td>
<td>+ REM large</td>
<td>0.989</td>
<td>0.925</td>
<td>3.323</td>
</tr>
<tr>
<td>WoW</td>
<td>+ REM base</td>
<td>0.998</td>
<td>0.938</td>
<td>12.742</td>
</tr>
<tr>
<td>WoW</td>
<td>+ REM large</td>
<td>0.997</td>
<td>0.913</td>
<td>10.788</td>
</tr>
<tr>
<td>CMUDoG</td>
<td>+ REM base</td>
<td>1.000</td>
<td>1.000</td>
<td>29.775</td>
</tr>
<tr>
<td>CMUDoG</td>
<td>+ REM large</td>
<td>0.999</td>
<td>0.991</td>
<td>27.247</td>
</tr>
</tbody>
</table>

Table 7: The result on the validation set used for training and evaluating the models with both tasks, NER and GEN. Larger models show slightly higher scores in GEN, but lower scores in ACC.

D Threshold

To utilize the off-the-shelf metric as our scoring function for filtering, we manually find the threshold both for effectiveness and efficiency. The high threshold is intended for evaluating the source-faithfulness of a given utterance strictly. A low threshold, on the other hand, targets only those utterances that are not faithful to the knowledge and aims to regenerate only fatal cases. As shown in Figure 6, we use a threshold of 0.5 to consider all datasets and 1.0 if we need to refine all utterances.
Figure 6: Manual search for the threshold of filtering module. DAE score in the y-axis indicates the source-faithfulness score after refining the utterances under the threshold.

E Prompts for Adversarial Data Generation with ChatGPT

The prompt used for generating adversarial data with ChatGPT is in Table 8:

Table 8: The prompt used for generating adversarial data.

F Refining Utterances of Larger Models

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Source-faithfulness</th>
<th>Ref. Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EC</td>
<td>TC</td>
</tr>
<tr>
<td>FoCus</td>
<td>$LLM_{KGC}$</td>
<td>42.74</td>
<td>26.72</td>
</tr>
<tr>
<td></td>
<td>+ BART$<em>{REM</em>{large}}$</td>
<td>46.06</td>
<td>28.76</td>
</tr>
<tr>
<td>WoW</td>
<td>$LLM_{KGC}$</td>
<td>6.91</td>
<td>5.87</td>
</tr>
<tr>
<td></td>
<td>+ BART$<em>{REM</em>{large}}$</td>
<td>15.97</td>
<td>11.15</td>
</tr>
<tr>
<td>CMUDoG</td>
<td>$LLM_{KGC}$</td>
<td>4.77</td>
<td>7.24</td>
</tr>
<tr>
<td></td>
<td>+ BART$<em>{REM</em>{large}}$</td>
<td>6.27</td>
<td>9.63</td>
</tr>
</tbody>
</table>

Table 9: The result of BART$_{REM_{large}}$ (denoted with +) refining the KGC utterances of large language model ($LLM_{KGC}$). The utterances are filtered with $\tau = 0.5$.

G Human Evaluation Questionnaire

<table>
<thead>
<tr>
<th>Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>The given sentences from model A and model B are the generated result of each machine consulting its knowledge to produce informative utterances. Considering the sentences produced by model A and model B, name the model that did better on the following three criteria.</td>
</tr>
<tr>
<td>Fluent: The response is linguistically fluent and not awkward.</td>
</tr>
<tr>
<td>Factually correct: The response is consistent with the given knowledge and does not generate the facts that are not given.</td>
</tr>
<tr>
<td>Well-paraphrased: The response is well-paraphrased rather than just copying and pasting the given knowledge.</td>
</tr>
</tbody>
</table>

Table 10: An example of the questionnaire for human evaluation.

H Prompts for LLM and LLM$_{REM}$

The prompts for refining utterances with LLM and LLM$_{REM}$ are in Table 11.

Table 11: The prompts for refining utterance by LLM and LLM$_{REM}$.
I REM Generation Examples

<table>
<thead>
<tr>
<th>[Knowledge]</th>
<th>The Václav Havel Library, located in Prague, is a charitable organization founded by Dagmar Havlová, Karel Schwarzenberg and Miloslav Petrusék on 26 July 2004.</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ REMbase : Václav Havel Library was founded on 26 July 2004 by Dagmar Havlová, Karel Schwarzenberg and Miloslav Petrusék.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[Knowledge]</th>
<th>The design of the house is attributed to John Verge and John Bibb and was built from 1835 to 1839 by James Hume.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Persona Knowledge]</td>
<td>I love colonial architecture. I like Australia. I would like to visit a museum. I would love to see a grotto. I have read about the Cadigal people.</td>
</tr>
<tr>
<td>[Utterance]</td>
<td>BARTbase: was from 35 18 to 39</td>
</tr>
<tr>
<td>+ REMbase : it was built from 1835 to 1839 by James Hume. I got this information on the internet. It was designed by John Verge and John Bibb.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[Knowledge]</th>
<th>A tattoo is a form of body modification where a design is made by inserting ink, dyes and pigments, either indelible or temporary, into the dermis layer of the skin to change the pigment.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Utterance]</td>
<td>BARTbase: I’m not sure if it is really a tattoo artist, but I do know that</td>
</tr>
<tr>
<td>+ REMbase : of course, it is a form of body modification where a design is made by inserting ink, dyes and pigments into the dermis layer of the skin to change the pigment.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[Knowledge]</th>
<th>It was also applied to the pit-style cooking techniques now frequently used in the Southeastern United States.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Utterance]</td>
<td>BARTbase: I’ll have to try it sometime! I’ve heard it’s used in the</td>
</tr>
<tr>
<td>+ REMbase : I don’t know if I’d ever try it. I know that pit-style cooking techniques are now frequently used in the Southeastern United States.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[Knowledge]</th>
<th>Leonardo DiCaprio as Dom Cobb, a professional thief who specializes in conning secrets from his victims by infiltrating their dreams. Joseph Gordon-Levitt as Arthur, Cobb’s partner who manages and researches the missions. ... Tom Hardy as Eames, a sharp-tongued associate of Cobb. wildly ingenious chess game, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Utterance]</td>
<td>BARTbase: have you seen the movie Inception?</td>
</tr>
<tr>
<td>+ REMbase : I’m good, how are you? Have you seen Inception? It’s a science fiction movie starring Leonardo DiCaprio, Joseph Gordon-Levitt, and Tom Hardy.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[Knowledge]</th>
<th>Angelina Jolie as Maleficent Sharlto Copley as Stefan Elle Fanning as Aurora Sam Riley as Diaval Imelda Staunton as Knotgrass Maleficent’ demolishes stereotypes ... Robert Stromberg SuperHero Maleficent is a 2014 American dark fantasy film directed by Robert Stromberg from a screenplay by Linda Woolverton, and starring Angelina Jolie as Maleficent with Sharlto Copley, Elle Fanning, Sam Riley, Imelda Staunton, Juno Temple, and Lesley Manville in supporting roles. ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Utterance]</td>
<td>BARTbase: have you seen Maleficent before?</td>
</tr>
<tr>
<td>+ REMbase : movie we’re talking about is Maleficent. Have you seen it? It’s a fantasy from 2014. It stars Angelina Jolie, Sharlto Copley, Elle Fanning, Sam Riley, and Imelda Staunton.</td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Examples of the fine-tuned baseline utterances refining by REMbase on KGC datasets.
FoCus

[Knowledge]
Elizabeth Bay House is a heritage-listed Colonial Regency style house and now a museum and grotto, located at 7 Onslow Avenue in the inner eastern Sydney suburb of Elizabeth Bay in the City of Sydney local government area of New South Wales, Australia.

[Persona Knowledge]
I love colonial architecture. I like Australia. I would like to visit a museum. I would love to see a grotto. I have read about the Cadigal people.

[Utterance]
BART_large : Elizabeth House located Australia which like. is place you like
+ REM_large : This place is called Elizabeth Bay House and it is located in Australia, which you like. It is a heritage-listed Colonial Regency style house and now a museum and grotto.

[Knowledge]
The Memphis Zoo, located in Midtown, Memphis, Tennessee, United States, is home to more than 3,500 animals representing over 500 different species.

[Persona Knowledge]
I hope to visit Tennessee. I like to ride on a tram. I have never seen a wetland. I would like to see a Giant Panda. I wish I could see a penguin.

[Utterance]
BART_large : is Memphis in. is place hope visit.
+ REM_large : It’s the Memphis Zoo in Tennessee, a place you hope to visit. It is home to more than 3,500 animals, 500 different species.

WoW

[Knowledge]
Competing theories state that Hindu identity developed in the British colonial era, or that it developed post-8th century CE after the Islamic invasion and medieval Hindu-Muslim wars.

[Utterance]
BART_large : I’m not fluent in English, but I understand that Hindu identity developed in the
+ REM_large : No, I’m not fluent in English, but I understand that Hindu identity developed in the British colonial era, or that it developed post-8th century CE.

CMUDoG

[Knowledge]
Sally Hawkins as Elisa Esposito, a mute cleaner who works at a secret government laboratory. ... The Shape of Water Rotten Tomatoes: 92% and average: 8.4/10 Metacritic Score: 87/100 CinemaScore: A 2017

[Utterance]
BART_large : have you seen The Shape of Water?
+ REM_large : Hello, Have you seen The Shape of Water? It’s a really good movie. It has a 92% on Rotten Tomatoes and a Metacritic Score of 87/100.

[Knowledge]
Lindsay Lohan as Cady Heron Rachel McAdams as Regina George Lacey Chabert as Gretchen Wieners Amanda Seyfried as Karen Smith Lizzy Caplan as Janis Ian Boasts a one-two-three punch in star Lindsay Lohan, screenwriter Tina Fey and director Mark Waters, and, indeed, it delivers a knockout The screenplay was highly praised by critics with Peter Travers of Rolling Stone calling it ‘comic gold’.

[Utterance]
BART_large : what did you think of it?
+ REM_large : I really liked it, I thought Lindsay Lohan did a great job as Cady Heron. Tina Fey did an amazing job writing the screenplay as well.

Table 13: Examples of the fine-tuned baseline utterances refining by REM_large on KGC datasets.
Focus

Knowledge
hesselberg (pronounced [ ] (about this sound:listen); 689 m above sea level) is the highest point in middle franconia and the franconian jura and is situated 60 km south west of nuremberg, germany.

Persona Knowledge
I love to enjoy nature. I live near Nuremberg. I am from Germany. I like heights. I study geology.

Utterance
INFO: The name of this place is Hesselberg.
+ REMlarge: The name of this place is Hesselberg, and it is located 60 km south west of nuremberg, germany, the country you are from.

Knowledge
st. george’s united methodist church, located at the corner of 4th and new streets, in the old city neighborhood of philadelphia, is the oldest methodist church in continuous use in the united states, beginning in 1769.

Persona Knowledge
I will stay in Philadelphia. I am going to Pennsylvania. I like history. I am religious. I am interested in the American Revolution.

Utterance
INFO: This is St. George’s United Methodist Church you may know of it because you are religious.
+ REMlarge: This is St. George’s United Methodist Church, you may know of it because you are religious. It is the oldest methodist church in continuous use in the united states.

Wow(Kilt)

Knowledge
red ketchup is a cult Quebec comic book series featuring FBI’s crazed rogue agent, Steve "Red" Ketchup. The series was created by Pierre Fournier and Réal Godbout, and will soon be adapted into a live action feature film by Martin Villeneuve for GO Films in Montreal.

Utterance
EDMem: ketchup red is a sweet sauce made from tomatoes.
+ REMlarge: I love red ketchup, which is a cult Quebec comic book series featuring FBI’s crazed rogue agent, Steve "Red" Ketchup. The series was created by Pierre Fournier and Réal Godbout.

Knowledge
childhood obesity is a condition where excess body fat negatively affects a child’s health or well-being. As methods to determine body fat directly are difficult, the diagnosis of obesity is often based on BMI. Due to the rising prevalence of obesity in children and its many adverse health effects it is being recognized as a serious public health concern. The term overweight rather than obese is often used when discussing childhood obesity, especially in open discussion, as it is less stigmatizing.

Utterance
EDMem: i’m not sure, but i do know that in the united states, overweight people are more likely to be obese than healthy people.
+ REMlarge: Yea i’m not sure, but the prevalence of obesity in children and its many adverse health effects it is being recognized as a serious public health concern.

CmuDnG

Knowledge
Leonardo DiCaprio as Frank Abagnale, Jr. Tom Hanks as Carl Hanratty Christopher Walken as Frank Abagnale, Sr. Nathalie Baye as Paula Abagnale. ... Catch me if you can Rotten Tomatoes: 96% and average: 7.9/10 Metacritic Score: 76/100 CinemaScore: A- 2002

Utterance
ITDD: yes, it has a 96 percent rating on rotten tomatoes and 7.9/10 on metacritic and a cinema score of an A- which is pretty good.
+ REMlarge: Yes, they do.

Knowledge
Mike and Sulley are led away by the CDA, and are expelled from the university as a result of their actions, but the other members of Oozma Kappa are accepted into the scare program the next semester, as Hardscrabble was impressed by their performances in the Games. As Mike leaves on the bus, Sulley runs after him to raise his spirits. Hardscrabble then appears and wishes the two luck, claiming they were the first students to have surprised her. The two take jobs in the mail room of Monsters, Inc., eventually working their way up to the scarer team.

Utterance
ITDD: Yes, they do. Sulley is a good friend of Mike’s and tries to cheer him on when he leaves on the bus to try to raise his spirits.

Table 14: Examples of the existing model utterances refining by REMlarge on KGC datasets.