Contrastive Learning Reduces Hallucination in Conversations

Weiwei Sun[1], Zhengliang Shi[1], Shen Gao[1], Pengjie Ren[1], Maarten de Rijke[2], Zhaochun Ren[1*]

[1]Shandong University, Qingdao, China
[2]University of Amsterdam, Amsterdam, The Netherlands
{weiwei.sun,shizhl}@mail.sdu.edu.cn, {shengao,renpengjie,zhaochun.ren}@sdu.edu.cn, m.derijke@uva.nl

Abstract

Pre-trained language models (LMs) store knowledge in their parameters and can generate informative responses when used in conversational systems. However, LMs suffer from the problem of "hallucination:" they may generate plausible-looking statements that are irrelevant or factually incorrect. To address this problem, we propose a contrastive learning scheme, named MixCL. A novel mixed contrastive objective is proposed to explicitly optimize the implicit knowledge elicitation process of LMs, and thus reduce their hallucination in conversations. We also examine negative sampling strategies of retrieved hard negatives and model-generated negatives. We conduct experiments on Wizard-of-Wikipedia, a public, open-domain knowledge-grounded dialogue benchmark, and assess the effectiveness of MixCL. MixCL effectively reduces the hallucination of LMs in conversations and achieves the highest performance among LM-based dialogue agents in terms of relevancy and factuality. We show that MixCL achieves comparable performance to state-of-the-art KB-based approaches while enjoying notable advantages in terms of efficiency and scalability.

1 Introduction

Open-domain dialogue agents have received increasing attention in recent years (Freitas et al. 2020; Huang, Zhu, and Gao 2020). In an engaging open-domain dialogue, a large amount of knowledge, such as commonsense (Young et al. 2018) and factual knowledge (Dinan et al. 2019), is involved. To integrate knowledge into dialogue agents, *KB-based methods* have been proposed to explicitly acquire knowledge from knowledge bases (Young et al. 2018; Dinan et al. 2019). However, KB-based methods suffer from problems of retrieval error (Liu et al. 2022) and inefficiency (Xu et al. 2022). Meanwhile, recent years have witnessed a rapid development of pre-trained language models (LMs) (Devlin et al. 2019; Brown et al. 2020) and their applications to dialogue tasks (Thoppilan et al. 2022). Large LMs implicitly store knowledge in their parameters during the pretraining stage (Petroni et al. 2019; Zhou et al. 2020) and thus, to some extent, they can serve as knowledge bases to ground open-domain dialogues (Zhao, Wu, and Xu 2020). Such approaches, known as *LM-based methods*, achieve promising performance in generating informative responses and obviate the drawbacks of KB-based methods. However, LM-based methods have the problem of "hallucination" (Shuster et al. 2021; Ji et al. 2022): they generate plausible-looking statements that are irrelevant or factually incorrect.

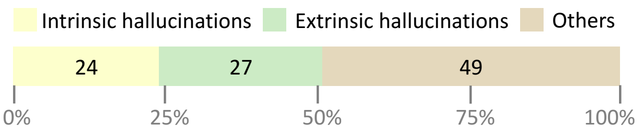


Figure 1: Results of a pilot experiment where annotators were asked to label 200 responses generated by BART on the Wizard-of-Wikipedia dataset for hallucination.

To understand the severity of hallucinations of LMs, we conduct a pilot experiment. We sample 200 responses generated by BART (Lewis et al. 2020) on the Wizard-of-Wikipedia dataset (Dinan et al. 2019) for various topics and conversation turns. These responses are annotated by three well-informed experts in terms of knowledge relevancy and factuality. Based on the results, we group the hallucinations of LMs into two types: *intrinsic hallucinations* and *extrinsic hallucinations*. Intrinsic hallucinations are non-factual statements, such as incorrectly predicting a celebrity's birthday. Extrinsic hallucinations are irrelevant or out-of-context responses, such as the a description of the history of football when the user asks the number of teams currently in the NFL. Fig. 1 summarizes the outcomes: intrinsic and extrinsic hallucinations account for 24% and 27% of the responses, respectively.

The problem of hallucinations is mainly attributable to the optimization recipes: the commonly used maximum likelihood estimation (MLE) with teacher forcing training encourages the model to imitate the training data blindly, leading to model hallucinations at inference time (Kang and Hashimoto 2020). Most studies on tackling hallucination in conversations focus on KB-based methods and use pre-retrieval (Shuster et al. 2021) or post-editing techniques (Dziri et al. 2021) to improve faithfulness; the hallucination of LM-based agents in eliciting knowledge

*Corresponding author.

inside LMs' parameters is still underexplored.

In this paper, we propose **Mix**ed **C**ontrastive **L**earning (MixCL) to alleviate the hallucinations of LM-based dialogue agents. MixCL explicitly samples the most confusing knowledge to the model and reduces its generation probability by contrasting it with the ground-truth. To this end, two novel steps are used by MixCL: (i) negative sampling, and (ii) mixed-contrastive learning. In the former, we sample the most confused negative knowledge by retrieving from the corpus or deriving via model bootstrapping. In the latter, we propose mixed-contrastive learning under the inspiration of mix-up data augmentation (Zhang et al. 2018), which mixes the positive and negative at span level. Moreover, we propose two mixed strategies regarding the two types of hallucination: entity-based mix-up and constituency-based mix-up. Finally, MixCL is optimized in an end-to-end manner, thus avoiding the retrieval step during inference and instead using the knowledge inside its parameters.

We conduct experiments on Wizard-of-Wikipedia (Dinan et al. 2019), an open-domain, knowledge-grounded dialogue dataset. Extensive experiments show that MixCL improves the informativeness and relevancy of the responses. Compared with previous LM-based methods (Zhao, Wu, and Xu 2020; Xu et al. 2022; Liu et al. 2022), MixCL achieves improvements by 5% to 15% in terms of response quality and relevancy. Moreover, MixCL achieves comparable performance as state-of-the-art KB-based methods (e.g., KnowledGPT (Zhao et al. 2020)), while speeding up 5× in model inference and showing superior scalability. The effectiveness of MixCL is also verified through human evaluation and ablation experiments.

Our contributions are as follows: (i) We propose MixCL, which reduces hallucinations of LMs in conversation through contrastive learning. (ii) We propose a hard negative sampling strategy to obtain the most confused negative knowledge (see Section 5.1). (iii) We propose a mix contrastive objective to optimize the model at span level (see Section 5.2). (iv) Experiments on the Wizard-of-Wikipedia dataset show that MixCL effectively reduces the hallucinating content produced by the LM and achieves comparable performance to KB-based approaches.[1]

2 Related Work

2.1 Knowledge-Grounded Dialogues

In open-domain knowledge-grounded dialogues (KGDs), people respond to each other's utterances in a meaningful way by integrating knowledge (Young et al. 2018; Huang, Zhu, and Gao 2020). To integrate knowledge, *KB-based methods* have been explored (Liu et al. 2018; Young et al. 2018; Dinan et al. 2019); they retrieve knowledge from a corpus through additional information retrieval (IR) modules. Studies on KB-based methods focus on knowledge selection (Meng et al. 2020; Shuster et al. 2021) and knowledge-grounded response generation (Zhao et al. 2020; Zheng and Huang 2021). However, KB-based methods suffer from the problems of retrieval errors (Liu et al. 2022), inefficiencies (Xu et al. 2022), and multi-granularity knowledge integration (Wu et al. 2022).

[1]We release our code at https://github.com/sunnweiwei/MixCL.

2.2 Language Models as Knowledge Bases

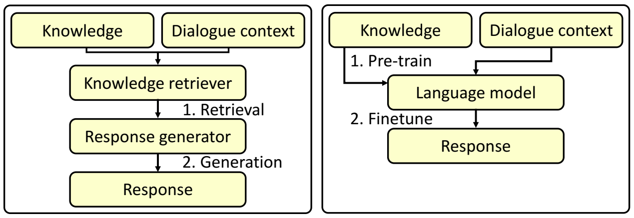
Recent years have witnessed a rapid development of language models (LMs) (Brown et al. 2020) and LM-based dialogue agents (Thoppilan et al. 2022). Large LMs store knowledge into their parameters during pre-training and can generate informative responses in conversations (Zhao, Wu, and Xu 2020). Petroni et al. (2019) show that LMs can serve as knowledge bases for downstream tasks (e.g., question answering (Roberts, Raffel, and Shazeer 2020)). On this basis, Zhao, Wu, and Xu (2020) show that LMs can ground open-domain dialogues using their implicit knowledge. Madotto et al. (2020) embed knowledge bases into model's parameters for end-to-end task-oriented dialogues. Roller et al. (2021) finetune LMs on KGD data. Cui et al. (2021) propose knowledge-enhanced finetuning methods to handle unseen entities. Xu et al. (2022) propose a topic-aware adapter to adapt LMs in KGDs. Liu et al. (2022) propose a multi-stage prompting approach for triggering knowledge in LMs. Wu et al. (2022) propose lexical knowledge internalization to integrate token-level knowledge into the model's parameters. However, existing LM-based methods suffer from the problem of hallucination. In this paper, we optimize the implicit knowledge eliciting process, i.e., reduce hallucination of LMs in KGD, via the proposed contrastive learning framework MixCL.

2.3 Contrastive Learning

Contrastive learning (CL) (Chopra, Hadsell, and LeCun 2005; Chen et al. 2020b) is based on the idea that similar samples should also be close in representation space, and has seen applications in NLP (Gao, Yao, and Chen 2021). CL has been used for optimizing knowledge retrieval processes (Karpukhin et al. 2021; Xiong et al. 2021), where the model learns to identify positive knowledge from negatives. On the task of neural text generation, CL (Jiang et al. 2022), a.k.a. unlikelihood training (Welleck et al. 2020) or negative training (He and Glass 2020), alleviates undesirable properties of the generated output, e.g., repetition (Shirai et al. 2020; Jiang et al. 2022), maliciousness (He and Glass 2020), dullness (Li et al. 2020b, 2022), or inconsistency (Li et al. 2020a). Moreover, Cao and Wang (2021) propose a sentence level contrastive learning method to reduce the hallucinations of text summarization model. Unlike existing studies, we propose a mixed contrastive learning framework MixCL that eliminates the hallucination at the span level with effective negative sampling strategies.

3 Problem Formulation

Let x, y, and k be the dialogue context, the corresponding response, and the ground-truth knowledge, respectively. As illustrated in Fig. 2, given a knowledge corpus $\mathcal{K}$, a dialogue agent learns to predict an informative response y based on the dialogue context x using the knowledge in $\mathcal{K}$. As



(a) KB-based dialogue agents explicitly retrieve text-based knowledge from corpus. (b) LM-based dialogue agents store knowledge in LM parameters and generate responses using implicit knowledge.

Figure 2: Types of dialogue agents.

discussed earlier, two approaches are studied in KGD, *KB-based methods* and *LM-based methods*. In this paper, we focus on the latter one.

KB-based Methods. KB-based dialogue agents (Dinan et al. 2019) ground the response generation by explicitly retrieving knowledge from $\mathcal{K}$. Two sub-modules, i.e., knowledge retriever and response generator, are employed by KB-based approaches, as shown in Fig. 2 (a).

LM-based Methods. In this paper, we explore language models as knowledge bases for dialogue agents (Zhao, Wu, and Xu 2020; Xu et al. 2022), as illustrated in Fig. 2 (b). In LM-based approaches, the LMs are first pre-trained on $\mathcal{K}$ to store the knowledge in their parameters. Then, the models directly generate y given x using the knowledge in their parameters and getting rid of the explicit retrieval step.

4 Preliminaries

We propose a LM-based dialogue agent for open-domain KGD. The proposed model $p_\theta(y|x)$ is based on a transformer-based language model with encoder-decoder architecture. The model is first pre-trained on the corpus $\mathcal{K}$ and then finetuned on dialogue data to generate informative responses.

Pre-training on Knowledge Corpus. We employ BART (Lewis et al. 2020) as the pre-trained transformer, which is pre-trained by denoising self-supervised learning:

$$\mathcal{L}_{\text{LM}} = -\mathbb{E}_{k\sim\mathcal{K}} \log p_\theta(k|\hat{k}), \tag{1}$$

where $\mathcal{K}$ is a text-based knowledge corpus (e.g., Wikipedia), k is a text sampled from knowledge corpus $\mathcal{K}$, and $\hat{k}$ denotes corrupted text by corruption functions (e.g., masking, deletion, infilling, etc.; Lewis et al. (2020)).

Finetuning on Dialogue Datasets. With the pre-trained LM, the model generates the response y given x without explicit knowledge retrieval step (Zhao, Wu, and Xu 2020; Xu et al. 2022). Maximum likelihood estimation (MLE) training loss on dialogue data with paired (x, y) is employed by previous methods. In MLE, the model learns to predict the ground-truth tokens for each step in a teacher forcing paradigm (Zhao, Wu, and Xu 2020; Xu et al. 2022):

$$\mathcal{L}_{\text{MLE}} = -\log p_\theta(y|x) = -\sum_{t=1}^{|y|} \log p_\theta(y_t|y_{<t}, x). \tag{2}$$

However, despite its effectiveness in generating informative responses, MLE loss encourages the model to imitate the training data blindly and leads to model hallucination (Kang and Hashimoto 2020). Studies have found that models trained with standard MLE may over-rely on previously predicted tokens, exacerbating error propagation (Wang and Sennrich 2020). As a result, during the inference stage, as the generated sequence grows, the errors accumulate along the sequence, and the model tends to amplify errors and generate hallucinating contents. We propose a novel contrastive learning framework MixCL to address this problem.

5 MixCL

Next, we present the proposed MixCL framework for addressing the hallucination of LMs. MixCL explicitly samples negative knowledge (i.e., non-factual or irrelevant knowledge) and reduces the generation probability of negative tokens by LMs through contrastive learning. As illustrated in Fig. 3, MixCL consists of two steps: *negative sampling* and *mixed contrastive learning*. In this section, we first present the negative sampling methods, then the mixed contrastive learning, and finally our optimization strategies.

5.1 Negatives Sampling

We sample negative knowledge for the dialogue context to construct training examples for contrastive learning. Formally, let z^+ be positive knowledge, i.e., a factual and relevant knowledge snippet, and let $\mathcal{Q}_{\text{Pos}}(x)$ be the collection of positive knowledge regarding x, where the $z^+ \sim \mathcal{Q}_{\text{Pos}}(x)$ is sampled from it. Here, $\mathcal{Q}_{\text{Pos}}(x)$ can be obtained through human labeling (Dinan et al. 2019) or heuristic methods (Zhao et al. 2020). We define z^- as negative knowledge, i.e., a non-factual or irrelevant knowledge snippet for x. Then, negative sampling is applied to construct the snippets z^- where the model is most likely to get confused. We introduce two methods for negative sampling, i.e., *retrieved negatives* and *model-generated negatives*, as illustrated in Fig. 3.

Retrieved Negatives. For a given x, a retrieval tool $\text{Ret}(*)$ is employed to retrieve irrelevant but potentially confusing knowledge from knowledge corpus $\mathcal{K}$:

$$\mathcal{Q}_{\text{Ret}}(x) = \{z^-|z^- \in \text{Ret}(x, \mathcal{K}), z^- \notin \mathcal{Q}_{\text{Pos}}(x)\}, \tag{3}$$

where $\text{Ret}(\cdot,\cdot)$ is implemented as TF-IDF retriever (Dinan et al. 2019), and $z^- \notin \mathcal{Q}_{\text{Pos}}(x)$ imposes the constraint that negative knowledge snippets should not be included in the positive knowledge.

Model-Generated Negatives. We also exploit a model bootstrapping approach, in which we generate knowledge by a model $p_\theta(z|x)$ and retain the examples where hallucination exist. We define:

$$\mathcal{Q}_{\text{Model}}(x) = \{z^-|z^- \sim p_\theta(z|x), z^- \cap \mathcal{Q}_{\text{Pos}}(x) = \varnothing\}, \tag{4}$$

where $z^- \sim p_\theta(z|x)$ denotes a negative knowledge snippet sampled from the LM with θ, and $z^- \cup \mathcal{Q}_{\text{Pos}}(x) = \varnothing$ imposes the constraint that negative knowledge snippets

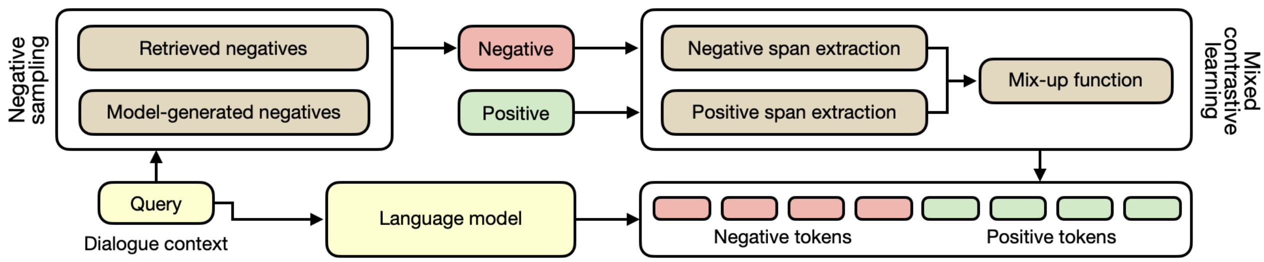


Figure 3: Overview of MixCL. MixCL consists of two steps: (i) negative sampling (Section 5.1), which samples most confusing negative knowledge to the model, and (ii) mixed contrastive learning (Section 5.2), which reduces the generation probability of negative tokens through mixed contrastive learning.

should not be included in the positive knowledge, which is implemented with a natural language inference (NLI) toolkit.[2]

On the basis of the above two methods, we define the constructed negative collection $\mathcal{Q}_{\text{Neg}}(x)$ with a hyper-parameter $\beta \in [0, 1]$ to control the relative contribution of the methods:

$$\mathcal{Q}_{\text{Neg}}(x) = \beta \mathcal{Q}_{\text{Ret}}(x) + (1-\beta)\mathcal{Q}_{\text{Model}}(x). \quad (5)$$

5.2 Mixed Contrastive Learning

Based on the positive knowledge z^+ and the sampled negative knowledge z^-, we introduce a contrastive learning framework to identify positive knowledge from negatives:

$$\mathcal{L}_{\text{CL}} = \mathbb{E}_{z^+ \sim \mathcal{Q}_{\text{Pos}}(x), \{z_i^-\}_{i=1}^M \overset{\text{iid}}{\sim} \mathcal{Z}} \, l(x, z^+, \{z_i^-\}_{i=1}^M, \theta), \quad (6)$$

where l denotes a contrastive loss function that is typically defined as cross-entropy loss l_{ce}[3] (Gao, Yao, and Chen 2021; Cao and Wang 2021), and M denotes the number of negatives.

However, l_{ce} only considers token-level or sentence-level contrast. It ignores fine-grained *span-level* contrast even though hallucinations typically exists at the span level. Therefore, inspired by work on mix-up data augmentation (Zhang et al. 2018; Kim et al. 2020; Shi, Livescu, and Gimpel 2021; Zhang, Yang, and Yang 2022), we propose a *mixed contrast objective*, which mixes the positive and negative examples into a sequence at the span level. As illustrated in Fig. 3, the proposed mixed contrastive learning method has three parts: (i) *extracting spans*, which extracts meaningful spans from both positive and negative knowledge; (ii) *mixing examples*, which mixes positive and negative knowledge using the extracted spans; and (iii) *mixed-contrast loss*, which optimizes the model at the span level through contrastive learning.

Extracting Spans. We extract the key components from the both positive and negative knowledge, z^+ and z^-. Regarding the two types of hallucinations, i.e., the intrinsic and extrinsic, we design two extraction strategies. As part of the pilot experiment reported in Section 1, we find that intrinsic hallucinations are typically associated with confused entities. Therefore, we use *named entity recognition* (NER)[4] to extract entities of various types e.g. `person` and `time`. Moreover, we find that extrinsic hallucination is mainly triggered by the emergence of irrelevant sentence fragments in the text. Therefore, we use *constituency parsing* (CP)[5] to extract sentence constituents, e.g., `noun` and `particle`. Through the two strategies, we extract sequence spans from z^+ and z^-, respectively.

Example. Consider knowledge snippets about the French soccer player Thierry Henry. A statement like *He was born and raised in Paris* would be in z^+, while the span "Montreal, Quebec, Canada" could be extracted from a snippet such as in *He was born in Montreal, Quebec, Canada* in z^-.

Mixing Examples. Based on the extracted spans, we mix the two examples z^+ and z^- into a mixed sequence $\tilde{z}$ via a mix-up function: $\tilde{z} = \text{Mix}(z^+, z^-)$. The mix-up function randomly selects a span in z^+, and then selects a span with the same type in z^- to substitute it. We define a sequence ϕ with the same length of $\tilde{z}$, which annotates the tokens in $\tilde{z}$ as 1 if they come from z^+ and 0 if they come from z^-.

In the earlier Thierry Henry example, the span "Paris" in a snippet in z^+ can be selected and substituted by the corresponding ones from a snippet in z^-, such as "Montreal, Quebec, Canada."

Mixed-Contrast Loss. Based on the mixed sequence $\tilde{z}$ and ϕ, we design a loss function l_{mix} as follows:

$$\begin{aligned} l_{mix}(z^+, z^-) = \\ & - \textstyle\sum_{j=1}^{|\tilde{z}_i|} [\, \phi_{i,j} \log p_\theta(\tilde{z}_{i,j} | \tilde{z}_{i,<j}, x) + \\ & \quad (1-\phi_{i,j}) \log(1 - p_\theta(\tilde{z}_{i,j} | \tilde{z}_{i,<j}, x))], \end{aligned} \quad (7)$$

where $\tilde{z}_i = \text{Mix}(z^+, z_i^-)$ is a mixed sequence of z^+ and z_i^-, and $\phi_{i,j}$ denotes the sign of token $\tilde{z}_{i,j}$, which equals 1 for positive tokens and 0 for negative tokens. Using the negative collection $\mathcal{Q}_{\text{Neg}}(x)$ defined in Eq. 5, we formalize the mixed contrast objective $\mathcal{L}_{\text{MCL}}$ as:

$$\textstyle\sum_{z^+ \sim \mathcal{Q}_{\text{Pos}}(x)} \sum_{z_i^- \sim \mathcal{Q}_{\text{Neg}}(x)}^{i=1,\ldots,M} l_{mix}(x, z^+, z_i^-, \theta). \quad (8)$$

[2] https://huggingface.co/roberta-large-mnli

[3] $l_{\text{ce}}(x, z^+, \{z_i^-\}_{i=1}^M, \theta) = -\log \frac{\exp p_\theta(z^+|x)}{\exp p_\theta(z^+|x) + \sum_{i=1}^M \exp p_\theta(z_i^-|x)}$.

[4] https://spacy.io/api/entityrecognizer/

[5] https://stanfordnlp.github.io/stanza/constituency.html

5.3 Optimization

During finetuning, MixCL is optimized by minimizing $\mathcal{L}_{\text{MCL}}$. Two additional loss are considered in training, i.e., $\mathcal{L}_{\text{LM}}$, $\mathcal{L}_{\text{MLE}}$. $\mathcal{L}_{\text{LM}}$ is used to alleviate catastrophic knowledge forgetting (Devlin et al. 2019; Chen et al. 2020a) and $\mathcal{L}_{\text{MLE}}$ is used to optimize the response generation ability. Therefore, the final training objective is defined as:

$$\mathcal{J}(\theta) = \alpha_1 \mathcal{L}_{\text{MLE}} + \alpha_2 \mathcal{L}_{\text{MCL}} + \alpha_3 \mathcal{L}_{\text{LM}}, \tag{9}$$

where three losses are optimized jointly and α_1, α_2, α_3 denote the weights of the three losses, respectively.

6 Experimental Setup

6.1 Datasets and Evaluation Metrics

We conduct experiments on the Wizard of Wikipedia (WoW) dataset. WoW is built with crowd-sourcing and employs Wikipedia as the knowledge corpus. WoW consists of 22,311 conversations over 1,365 general topics that range from e-books to toga parties to showers. The ground-truth knowledge used in each turn is manually labeled. The WoW test set is split into *test seen* and *test unseen* based on whether the topic appears in the training set. We evaluate our methods on both test seen and test unseen.

We choose F1, ROUGE, BLEU, MT, Knowledge-F1 (KF1), Entity-F1 (EF1), and Accuracy (Acc) as metrics. **F1** (Dinan et al. 2019) calculates the unigram F1 between the generated text and the ground-truth text. For **ROUGE** (Lin 2004) we use ROUGE-L (RL for short) following previous work. **BLEU** (Papineni et al. 2002) we use BLEU-2 and BLEU-4 (or B2 and B4 for short) and use the implementation in the NLTK Toolkit. **MT** (Meteor) (Denkowski and Lavie 2014) is based on the harmonic mean of unigram precision and recall. **Knowledge-F1** (Dinan et al. 2019) (or KF1 for short) calculates the F1 between the generated response and the ground-truth knowledge sentence, which indicates the informativeness of a response. **Acc** measures the knowledge selection accuracy. As we skip the knowledge selection step, we select knowledge by matching the generated response with each knowledge candidate in WoW using the F1 score. **Entity-F1** (or EF1 for short) identifies entities in text using Spacy, deletes the non-entity words, and calculates the F1 score between the modified generated text and the ground-truth response. EF1 eliminates the impact of the stop-word and focuses on the accuracy of entities.

In addition, we randomly sample 100 examples from the test seen and test unseen segments of the test set, respectively, and recruit three experts for human evaluation. Each annotator is presented with examples that come with dialogue context and model responses. Four metrics are considered in the human evaluation: **Informativeness**, which measures whether the response is knowledge-inclusive; **Relevancy**, which measures whether the response's content is relevant to the dialogue; **Factuality**, which measures whether the information in the response is factually correct;[6] and **Humanlikeness**, which measures whether the response is human-like in its fluency and naturalness. The annotators are asked to assign a score in $\{0, 1\}$ (representing "non-factual" and "factual") for factuality, and a score in $\{0, 1, 2\}$ (representing "bad" "fair", and "good") for the others.

[6] The human annotators used Google to check the factuality of the responses.

6.2 Baselines

We compare MixCL with baselines of two categories: (i) *KB-based methods* that use additional IR modules for explicit knowledge retrieval, and (ii) *LM-based methods* that use LMs as a knowledge base. All models are re-evaluated with the same evaluation function using the official public checkpoints.

The KB-based methods we consider are: **TMN** (Dinan et al. 2019) (50M), which combines a transformer with an external memory network to select knowledge and generate a response; **DukeNet** (Meng et al. 2020) (150M), which is the best performing KB-based method without using pre-trained LMs and which models knowledge shift with a dual learning scheme; **KnowledGPT** (Zhao et al. 2020) (227M), which exploits pre-trained LMs in a KB-based approach, selects knowledge using BERT, generates responses using GPT-2, and optimizes the two modules jointly with reinforcement learning; it achieves state-of-the-art performance. We also introduce **KnowBART** (600M), a KB-based model that selects knowledge using RoBERTa and generates responses using BART-Large.

The KB-based methods listed above retrieve knowledge under *oracle conditions*, i.e., they are given a small subset of Wikipedia with roughly ten passages that *definitely* contain the ground-truth knowledge (Dinan et al. 2019; Liu et al. 2022). We also consider KB-based methods under *realistic experimental conditions*, where passages from the full knowledge corpus (i.e., Wikipedia) are retrieved. We employ the state-of-the-art passage retrieval model GENRE (Cao et al. 2021) from the KILT leaderboard (Petroni et al. 2021), which is reported to outperform competitors (e.g., DPR and BM25) by a substantial margin on WoW.

The LM-based methods that we consider are: **GPT-2** (Zhao, Wu, and Xu 2020) (345M),which finetunes GPT-2 on knowledge-grounded dialogue data; **BlenderBot** (Roller et al. 2021) (400M), which pre-trains a transformer with encoder-decoder architecture on *reddit* data, and then finetunes the model on KGD data; **KnowExpert** (Xu et al. 2022) (117M), which uses a topic-aware adapter that first clusters Wikipedia using a topic model and then employs a mix-of-adapter architecture to adapt a GPT-2 model to open-domain dialogues; **MSDP** (Liu et al. 2022) (357M), which uses a multi-stage prompting model, designs task-specific prompts with task instructions and in-context examples, and uses Megatron-LM (Shoeybi et al. 2019) to produce knowledge and response in a two-stage process.

6.3 Implementation Details

We implement MixCL using BART-Large (400M) (Lewis et al. 2020) in HuggingFace's Transformers library. We use Wikipedia as the knowledge corpus $\mathcal{K}$, as it is used as knowledge corpus by WoW. We determine the hyperparameters through pilot experiments. We set the

Method	Test seen								Test unseen							
	F1	RL	B2	B4	MT	KF1	EF1	Acc	F1	RL	B2	B4	MT	KF1	EF1	Acc
KB-based methods under realistic conditions																
TMN (Dinan et al. 2019)	17.3	17.0	5.7	1.1	14.8	15.8	8.7	15.2	14.4	14.5	3.3	0.3	11.5	9.4	2.1	8.6
DukeNet (Meng et al. 2020)	18.5	17.7	6.4	1.9	16.0	18.5	12.0	20.6	15.9	15.9	4.8	1.1	13.7	14.7	8.0	14.3
KnowledGPT (Zhao et al. 2020)	21.1	20.1	8.9	3.4	20.0	22.2	15.5	24.3	19.5	18.4	8.0	2.6	18.3	20.0	11.7	20.2
KnowBART	21.1	18.9	8.5	3.3	17.8	21.3	16.2	24.2	21.0	18.3	8.9	3.6	17.9	22.5	16.2	24.0
KB-based methods under oracle conditions																
DukeNet (Meng et al. 2020)	19.3	18.7	7.5	2.5	17.2	19.6	13.2	22.1	17.1	17.0	6.0	1.7	15.2	16.5	9.2	16.8
KnowledGPT (Zhao et al. 2020)	22.0	20.8	9.9	3.7	20.9	23.8	16.9	26.3	20.5	19.5	8.7	3.0	19.3	22.1	13.3	22.6
KnowBART	22.1	19.6	9.1	3.7	18.1	23.1	18.0	26.8	22.7	20.1	9.8	4.3	18.7	24.1	18.4	27.5
LM-based methods																
GPT-2 (Zhao, Wu, and Xu 2020)	19.6	18.5	7.8	1.4	17.8	17.9	13.3	15.4	18.3	17.3	6.5	0.8	16.1	14.6	7.2	8.4
BlenderBot (Roller et al. 2021)	18.8	19.4	7.7	2.3	18.0	18.2	13.1	16.7	17.8	16.9	5.5	0.8	15.0	15.7	7.1	9.6
KnowExpert (Xu et al. 2022)	18.7	18.6	6.7	1.3	16.5	14.1	9.8	12.6	16.7	17.2	5.4	0.6	14.5	11.8	5.5	9.2
MSDP (Liu et al. 2022)	17.8	16.5	6.1	1.9	18.2	21.7	13.9	18.4	16.9	16.1	5.5	1.1	16.2	20.3	8.4	16.1
Ours	**21.6**	**20.5**	**9.2**	**2.7**	**20.5**	**22.3**	**16.3**	**20.4**	**19.6**	**18.8**	**7.4**	**1.4**	**18.0**	18.0	**11.6**	14.4

Table 1: Evaluation results on Wizard-of-Wikipedia. The first group lists *KB-based methods under realistic conditions*. The second group lists *KB-based methods under oracle conditions*. The third group lists *LM-based methods*, including MixCL. We highlight the results of MixCL that significantly exceed the previous-best LM-based methods in boldface (t-test, $p < 0.05$). We also highlight the best results of previous KB-based methods and LM-based methods by underlining them, respectively.

weight of the language model loss α_3 to 0.3 at initialization and linearly decay until 0. We set α_1 and α_2, i.e., the weight of the MLE loss and MCL loss, to 0.4 and 0.3, respectively, and linearly increase to 0.5 and 0.5. We use greedy decoding in testing. More details are available at https://github.com/sunnweiwei/MixCL.

7 Experimental Results

7.1 Results of Automatic Evaluation

Table 1 shows the results of automatic evaluation metrics. Overall, MixCL achieves the highest scores of the LM-based methods and competitive results compared to the KB-based methods under realistic conditions. Compared with previous LM-based methods (the third group in Table 1), MixCL achieves the highest scores on almost all metrics. For example, MixCL gets F1 $=$ 21.6, B4 $=$ 2.7 on test seen and F1 $=$ 19.6, B4 $=$ 1.4 on test unseen, with about 5% to 15% relative improvements over previous-best LM-based baselines. Moreover, we find a dilemma with the previous LM-based methods in terms of response quality (e.g., F1, RL, B2) and knowledge relevance (e.g., KF1, EF1, Acc). For example, MSDP performs well on knowledge relevance at the expense of response quality, while GPT-2 and BlenderBot show the opposite. MixCL, on the other hand, performs well on both fronts.

Furthermore, compared with KB-based methods (the first block in Table 1), we find that MixCL outperforms two non-LM methods (DukeNet and TMN) by a large margin. Compared to KnowledGPT and KnowBART, which combine LMs with the KB-based approach, MixCL outperforms them on test seen. On test unseen, MixCL lags behind the best performing KB-based baselines, probably due to knowledge forgetting issues.

Methods	Test seen				Test unseen			
	Info.	Rel.	Fact.	Hum.	Info.	Rel.	Fact.	Hum.
DukeNet^{K}	1.44	1.22	0.71	1.16	1.21	1.08	0.72	1.03
KnowledGPT^{K}	1.67	1.47	0.87	1.73	1.63	1.23	0.83	1.36
KnowBART^{K}	1.67	1.57	0.89	1.70	1.68	1.56	0.91	1.44
KnowExpert^{L}	1.45	1.36	0.62	1.45	1.49	1.26	0.59	1.15
MSDP^{L}	1.20	0.96	0.71	0.98	1.28	1.18	0.82	1.05
BART^{L}	1.51	1.45	0.76	1.58	1.50	1.47	0.82	1.40
Ours^{L}	1.71	1.55	0.89	1.77	1.67	1.53	0.87	1.47
Human	1.84	1.85	0.98	1.96	1.83	1.85	0.95	1.95

Table 2: Human evaluation results. Methods marked with ^{K} denote KB-based methods, and those marked with ^{L} denote LM-based methods. The four metrics (Info., Rel., Fact., and Hum.) denote informativeness, relevance, factuality, and humanlikeness, respectively.

Finally, under oracle conditions, the KB-based methods (the second group in Table 1) show better results than MixCL. However, the manually selected knowledge candidates include the ground-truth, which is unavailable in realistic scenarios.

7.2 Results of Human Evaluation

Table 2 shows the human evaluation results. The Fleiss' kappa value is above 0.60, indicating substantial agreement among the annotators. MixCL consistently outperforms LM-based baselines on all metrics, and also outperforms KB-based baselines in metrics. MixCL is capable of generating more informative responses compared to previous LM-based methods. Moreover, MixCL effectively

Methods	Test seen			Test unseen		
	F1	B4	KF1	F1	B4	KF1
Base model	21.6	2.7	22.3	19.6	1.4	18.0
-w/o $\mathcal{L}_{\text{MCL}}$	$21.0_{\downarrow 0.6}$	$2.0_{\downarrow 0.7}$	$19.1_{\downarrow 3.2}$	$19.1_{\downarrow 0.5}$	$1.0_{\downarrow 0.4}$	$16.9_{\downarrow 1.1}$
-w/o $\mathcal{Q}_{\text{Neg}}(*)$	$20.8_{\downarrow 0.8}$	$2.4_{\downarrow 0.3}$	$20.8_{\downarrow 1.5}$	$19.0_{\downarrow 0.6}$	$1.1_{\downarrow 0.3}$	$17.4_{\downarrow 0.6}$
-w/o $\mathcal{Q}_{\text{Model}}(*)$	$21.3_{\downarrow 0.3}$	$2.5_{\downarrow 0.2}$	$21.7_{\downarrow 0.6}$	$19.4_{\downarrow 0.2}$	$1.2_{\downarrow 0.2}$	$17.5_{\downarrow 0.5}$
-w/o $\mathcal{L}_{\text{LM}}$	$21.3_{\downarrow 0.3}$	$2.6_{\downarrow 0.1}$	$21.8_{\downarrow 0.5}$	$18.6_{\downarrow 1.0}$	$1.2_{\downarrow 0.2}$	$16.7_{\downarrow 1.3}$
-Only $\mathcal{L}_{\text{MLE}}$	$20.9_{\downarrow 0.7}$	$1.8_{\downarrow 0.9}$	$18.9_{\downarrow 3.4}$	$18.8_{\downarrow 0.8}$	$0.9_{\downarrow 0.5}$	$16.0_{\downarrow 2.0}$

Table 3: Ablation study. The base model, MixCL, is compared with several variants. See Section 7.3.

increases relevance and factuality, demonstrating its effectiveness in reducing both types of hallucinations. In particular, we find that KnowledGPT is outperformed by MixCL in terms of knowledge relevance, probably due to the presence of retrieval errors. Finally, MixCL's responses are considered more human-like by the annotators.

7.3 Ablation Studies

In Table 3, we compare MixCL with several ablative variants. The variants and our findings are as follows:

No $\mathcal{L}_{\text{MCL}}$ – We remove the mixed contrast objective. The performance of the model shows a notable degradation, especially for the knowledge relevance metric, i.e., KF1. This suggests that the proposed mixed contrast objective is effective in increasing the relevance of responses.

No $\mathcal{Q}_{\text{Neg}}(*)$ – We remove the hard negative sampling process and use randomly sampled instances as negatives. The effectiveness of the hard negative sampling is evidenced by the decrease in the metric on KF1.

No $\mathcal{Q}_{\text{Model}}$ – We remove the negatives generated by the model. The results suggest that model-generated negatives provides harder negative examples for the model, i.e., the knowledge that is more likely to be confounded by the LMs.

No $\mathcal{L}_{\text{LM}}$ – We remove the LM loss. The effect of the model is a decline, especially on unseen topics. This results suggests that LM loss is instrumental in suppressing the catastrophic knowledge forgetting problem of the LMs in conversations (Chen et al. 2020a).

Only $\mathcal{L}_{\text{MLE}}$ – This variant optimizes the model only by MLE loss. We observe a substantial performance drop, especially on KF1, which demonstrates the effectiveness of MixCL in improving the knowledge relevancy and factuality of LMs.

7.4 Efficiency Analysis

In Fig. 4, we compare MixCL against baselines in terms of efficiency and effectiveness. We adjust the inference efficiency of the models by evaluating the model with different numbers of parameters (e.g., 140M and 400M). Compared with KB-based methods, LM-based methods generally have an advantage in terms of speed as they get rid of the extra IR step. However, previous LM-based methods are outperformed by KB-based methods regarding response quality. By explicitly eliminating the hallucinations of LM in conversations, MixCL significantly improves the response quality of LM-based methods without compromising efficiency. Notably, MixCL is 5× more efficient than state-of-the-art KB-based methods while achieving competitive response generation performance. Moreover, the improvements of MixCL along with the model size are more noticeable compared to KB-based methods (see the dashed lines), indicating its superior ability to utilize the knowledge of pre-trained model.

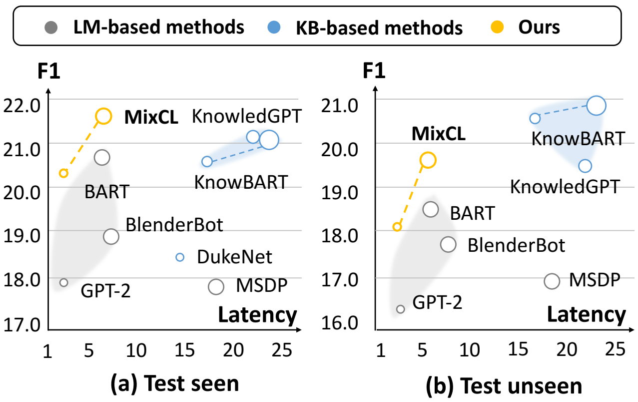


Figure 4: Latency (minutes) versus response quality (F1 score) on WoW test seen and test unseen. Gray, blue, and orange indicate LM-based, KB-based, and the proposed methods, respectively. The size of the circle indicates the number of parameters of these methods.

7.5 Case Study

We conduct several case studies and find that MixCL is more effective at incorporating knowledge and generating more engaging and human-like responses than baselines. Details about our case studies are available in https://github.com/sunnweiwei/MixCL.

8 Conclusions

In this paper, we have proposed MixCL, a contrastive learning framework aimed at reducing the hallucination of language models in conversations. MixCL is enhanced by negative sampling and mixed contrastive objective. Experiments on the Wizard-of-Wikipedia dataset have shown that MixCL outperforms existing LM-based methods and achieves comparable performance as state-of-the-art KB-based methods. Human evaluation and ablative experiments also confirm MixCL's effectiveness in eliminating hallucination of LMs. Moreover, MixCL demonstrates advantages in terms of efficiency and scalability. Hence, we believe that MixCL provides new insights on using knowledge inside large language models' parameters for KGD tasks.

The limitations of this work include the problem of knowledge forgetting. In future work, we would like to explore practical approaches to avoiding catastrophic knowledge forgetting. We also plan to reproduce our findings for other, less resource-rich languages.

Acknowledgements

This work was supported by the National Key R&D Program of China with grant No. 2020YFB1406704, the Natural Science Foundation of China (62272274, 62202271, 61902219, 61972234, 62072279, 62102234), the Natural Science Foundation of Shandong Province (ZR2021QF129), the Key Scientific and Technological Innovation Program of Shandong Province (2019JZZY010129), and by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T. J.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NIPS*, 1877–1901.

Cao, N. D.; Izacard, G.; Riedel, S.; and Petroni, F. 2021. Autoregressive Entity Retrieval. In *ICLR*.

Cao, S.; and Wang, L. 2021. CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization. ArXiv:abs/2109.09209.

Chen, S.; Hou, Y.; Cui, Y.; Che, W.; Liu, T.; and Yu, X. 2020a. Recall and Learn: Fine-tuning Deep Pretrained Language Models with Less Forgetting. In *EMNLP*, 7870–7881.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020b. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 1597–1607.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *CVPR*, 539–546 vol. 1.

Cui, L.; Wu, Y.; Liu, S.; and Zhang, Y. 2021. Knowledge Enhanced Fine-Tuning for Better Handling Unseen Entities in Dialogue Generation. In *EMNLP*, 2328–2337.

Denkowski, M. J.; and Lavie, A. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *WMT@ACL*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational agents. In *ICLR*.

Dziri, N.; Madotto, A.; Zaiane, O.; and Bose, A. 2021. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In *EMNLP*, 2197–2214.

Freitas, D. D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; and Le, Q. V. 2020. Towards a Human-like Open-Domain Chatbot. ArXiv:abs/2001.09977.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*, 6894–6910.

He, T.; and Glass, J. R. 2020. Negative Training for Neural Dialogue Response Generation. In *ACL*, 2044–2058.

Huang, M.; Zhu, X.; and Gao, J. 2020. Challenges in Building Intelligent Open-domain Dialog Systems. *ACM TOIS*, 38: 1 – 32.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; and Fung, P. 2022. Survey of Hallucination in Natural Language Generation. ArXiv:abs/2202.03629.

Jiang, S.; Zhang, R.; Vakulenko, S.; and de Rijke, M. 2022. A Simple Contrastive Learning Objective for Alleviating Neural Text Degeneration. ArXiv:abs/2205.02517.

Kang, D.; and Hashimoto, T. B. 2020. Improved Natural Language Generation via Loss Truncation. In *ACL*, 718–731.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L. Y.; Edunov, S.; Chen, D.; and tau Yih, W. 2021. Dense Passage Retrieval for Open-Domain Question Answering. In *NAACL*, 5835–5847.

Kim, S.; Lee, G.; Bae, S.; and Yun, S. 2020. MixCo: Mix-up Contrastive Learning for Visual Representation. In *NIPS | Workshop on Self-Supervised Learning*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, 7871–7880.

Li, M.; Roller, S.; Kulikov, I.; Welleck, S.; Boureau, Y.-L.; Cho, K.; and Weston, J. 2020a. Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training. In *ACL*, 4715–4728.

Li, X.; Li, P.; Wang, Y.; Liu, X.; and Lam, W. 2020b. Enhancing Dialogue Generation via Multi-Level Contrastive Learning. ArXiv:abs/2009.09147.

Li, Y.; Feng, S.; Sun, B.; and Li, K. 2022. Diversifying Neural Dialogue Generation via Negative Distillation. In *NAACL*, 407–418.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL*.

Liu, S.; Chen, H.; Ren, Z.; Feng, Y.; Liu, Q.; and Yin, D. 2018. Knowledge Diffusion for Neural Dialogue Generation. In *ACL*, 74–81.

Liu, Z.; Patwary, M. A.; Prenger, R. J.; Prabhumoye, S.; Ping, W.; Shoeybi, M.; and Catanzaro, B. 2022. Multi-Stage Prompting for Knowledgeable Dialogue Generation. In *Findings of ACL*, 1317–1337.

Madotto, A.; Cahyawijaya, S.; Winata, G. I.; Xu, Y.; Liu, Z.; Lin, Z.; and Fung, P. 2020. Learning Knowledge Bases with Parameters for Task-Oriented Dialogue Systems. In *Findings of ACL*.

Meng, C.; Ren, P.; Chen, Z.; Sun, W.; Ren, Z.; Tu, Z.; and de Rijke, M. 2020. DukeNet: A Dual Knowledge Interaction Network for Knowledge-Grounded Conversation. In *SIGIR*, 1151–1160.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*, 311–318.

Petroni, F.; Piktus, A.; Fan, A.; Lewis, P.; Yazdani, M.; Cao, N. D.; Thorne, J.; Jernite, Y.; Plachouras, V.; Rocktaschel, T.; and Riedel, S. 2021. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *NAACL*, 2523–2544.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language Models as Knowledge Bases? In *EMNLP*, 2463–2473.

Roberts, A.; Raffel, C.; and Shazeer, N. M. 2020. How Much Knowledge Can You Pack into the Parameters of a Language Model? In *EMNLP*, 5418–5426.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E. M.; Boureau, Y.-L.; and Weston, J. 2021. Recipes for Building an Open-Domain Chatbot. In *EACL*, 300–325.

Shi, H.; Livescu, K.; and Gimpel, K. 2021. Substructure Substitution: Structured Data Augmentation for NLP. In *Findings of ACL*, 3494–3508.

Shirai, K.; Hashimoto, K.; Eriguchi, A.; Ninomiya, T.; and Mori, S. 2020. Neural Text Generation with Artificial Negative Examples. ArXiv:abs/2012.14124.

Shoeybi, M.; Patwary, M. A.; Puri, R.; LeGresley, P.; Casper, J.; and Catanzaro, B. 2019. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. ArXiv:abs/1909.08053.

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of EMNLP*, 3784–3803.

Thoppilan, R.; Freitas, D. D.; Hall, J.; Shazeer, N. M.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; Li, Y.; Lee, H.; Zheng, H.; Ghafouri, A.; Menegali, M.; Huang, Y.; Krikun, M.; Lepikhin, D.; Qin, J.; Chen, D.; Xu, Y.; Chen, Z.; Roberts, A.; Bosma, M.; Zhou, Y.; Chang, C.-C.; Krivokon, I. A.; Rusch, W. J.; Pickett, M.; Meier-Hellstern, K. S.; Morris, M. R.; Doshi, T.; Santos, R. D.; Duke, T.; Søraker, J. H.; Zevenbergen, B.; Prabhakaran, V.; Diaz, M.; Hutchinson, B.; Olson, K.; Molina, A.; Hoffman-John, E.; Lee, J.; Aroyo, L.; Rajakumar, R.; Butryna, A.; Lamm, M.; Kuzmina, V. O.; Fenton, J.; Cohen, A.; Bernstein, R.; Kurzweil, R.; Aguera-Arcas, B.; Cui, C.; Croak, M.; Chi, E.; and Le, Q. 2022. LaMDA: Language Models for Dialog Applications. ArXiv:abs/2201.08239.

Wang, C.; and Sennrich, R. 2020. On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation. In *ACL*.

Welleck, S.; Kulikov, I.; Roller, S.; Dinan, E.; Cho, K.; and Weston, J. 2020. Neural Text Generation with Unlikelihood Training. In *ICLR*.

Wu, Z.; Bi, W.; Li, X.; Kong, L.; and Kao, B. C. 2022. Lexical Knowledge Internalization for Neural Dialog Generation. In *ACL*, 7945–7958.

Xiong, L.; Xiong, C.; Li, Y.; Tang, K.-F.; Liu, J.; Bennett, P.; Ahmed, J.; and Overwijk, A. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *ICLR*.

Xu, Y.; Ishii, E.; Liu, Z.; Winata, G. I.; Su, D.; Madotto, A.; and Fung, P. 2022. Retrieval-Free Knowledge-Grounded Dialogue Response Generation with Adapters. In *ACL | Workshop on DialDoc*, 93–107.

Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018. Augmenting End-to-End Dialog Systems with Commonsense Knowledge. In *AAAI*.

Zhang, H.; Cissé, M.; Dauphin, Y.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

Zhang, L.; Yang, Z.; and Yang, D. 2022. TreeMix: Compositional Constituency-based Data Augmentation for Natural Language Understanding. In *NAACL*, 5243–5258.

Zhao, X.; Wu, W.; Xu, C.; Tao, C.; Zhao, D.; and Yan, R. 2020. Knowledge-Grounded Dialogue Generation with Pre-trained Language Models. In *EMNLP*, 3377–3390.

Zhao, Y.; Wu, W.; and Xu, C. 2020. Are Pre-tr ained Language Models Knowledgeable to Ground Open Domain Dialogues? ArXiv:abs/2011.09708.

Zheng, C.; and Huang, M. 2021. Exploring Prompt-based Few-shot Learning for Grounded Dialog Generation. ArXiv:abs/2109.06513.

Zhou, X.; Zhang, Y.; Cui, L.; and Huang, D. 2020. Evaluating Commonsense in Pre-trained Language Models. In *AAAI*.