You Don't Know My Favorite Color: Preventing Dialogue Representations from Revealing Speakers' Private Personas

Haoran Li[1], **Yangqiu Song**[1], **Lixin Fan**[2]
[1]Dept. of CSE, Hong Kong University of Science and Technology
[2]AI Group, WeBank
hlibt@connect.ust.hk, yqsong@cse.ust.hk,
lixinfan@webank.com

Abstract

Social chatbots, also known as chit-chat chatbots, evolve rapidly with large pretrained language models. Despite the huge progress, privacy concerns have arisen recently: training data of large language models can be extracted via model inversion attacks. On the other hand, the datasets used for training chatbots contain many private conversations between two individuals. In this work, we further investigate the privacy leakage of the hidden states of chatbots trained by language modeling which has not been well studied yet. We show that speakers' personas can be inferred through a simple neural network with high accuracy. To this end, we propose effective defense objectives to protect persona leakage from hidden states. We conduct extensive experiments to demonstrate that our proposed defense objectives can greatly reduce the attack accuracy from 37.6% to 0.5%. Meanwhile, the proposed objectives preserve language models' powerful generation ability.

1 Introduction

Social chatbots have been widely used to benefit many applications from answering factual questions to showing emotional companionship. With recent progress in large pretrained language models (Radford et al., 2019; Yang et al., 2019), some attempts (Wolf et al., 2019; Zhang et al., 2020; Ham et al., 2020; Shen et al., 2021; Sevegnani et al., 2021; Gu et al., 2021b) are made to build chatbots based on large generative language models (LMs). To train such LM-based chatbots, private conversations are collected. Unfortunately, large language models tend to memorize training data and some private data can be recovered from models (Pan et al., 2020; Carlini et al., 2021). Besides such memorization problems, "overlearning" on simple training objectives can reveal sensitive attributes indirectly related to the learning task (Song and Shmatikov, 2020). LM-based social chatbots essentially inherit the privacy issues of general LMs and the overlearning problem.

For example, as Figure 1 shows, when using a fine-tuned GPT-2 as the encoder and decoder of an LM-based social chatbot, if the learned representation of each utterance can be obtained by an adversary, then the adversary can build a classifier to predict the persona information based on the representation. As shown by the example, for five out of 14 utterances, the attacker can successfully predict the persona, which can be harmful if the users (speakers of the utterances) do not prefer to reveal the persona information. Thus, in practice, when deploying such kinds of chatbots in real applications, we should first make sure that no private information can be leaked by the models.

To systematically study the privacy issues in LM-based social chatbots, there are several challenges. First, there is no existing data that can be used to quantify how much private information is revealed by an LM. Second, there has been no existing work showing how to attack utterance-level representations to obtain sensitive information. Third, there has been no existing LM-based chatbot that can defend against persona inference attacks, and no study shows how to protect both known and unknown persona attributes.

In this paper, to address the above challenges, we use the fine-tuned GPT-2 as our chatbot. We first collect a dataset by aligning personas with corresponding utterances in PersonaChat dataset (Zhang et al., 2018). Then we show that "overlearning" can happen for LM-based chatbots to reveal personas of speakers. We build a single external multi-layer perception (MLP) attacker model to perform black-box persona inference attacks on the utterance-level embeddings. With no access to parameters of the chatbot, the attacker model can infer speakers' personas with 37.59% accuracy over 4,332 personas. The high accuracy of the attacker model implies that the utterance-level embeddings have potential vulnerabilities to reveal

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5858 - 5870
July 10-15, 2022

	Context	Attacks on LM		Attacks on the defensed LM	
Human A	Hello, how are you tonight?	I take things very literally	✗	I am engaged to be married	✗
Human B	Hello my friend. I am well.	I am a happy person	✗	I like to go shopping with my daughters	✗
Human A	Good, glad to hear it. What do you do for fun?	I do whatever it takes to get what I want	✗	My favorite color is blue	✗
Human B	I ride around the town on my cool bicycle.	I love to ride my bike on the weekend	✗	My favorite color is blue	✗
Human A	Really? I really like mountain bike too.	I also like to mountain bike	✓	My favorite color is blue	✗
Human B	I wish I lived in the mountains.	I have never been out of the country	✗	My favorite color is blue	✗
Human A	Do you like nature? I have been to 12 national parks.	I like to visit national parks	✓	My favorite color is blue	✗
Human B	I love nature. I like looking at plants.	I really love plants	✓	My favorite color is blue	✗
Human A	I love plants too, and hiking. In fact, I am actually an environmental activist.	I am an environmental engineer	✓	My favorite color is blue	✗
Human B	Cool, I am a vegan.	I am a vegan	✓	My favorite color is blue	✗
Human A	Nice, do you have a favorite food?	I love ham and cheese sandwiches	✗	I have my own salon	✗
Human B	My favorite dish is lentil curry.	My favorite meal is chicken and rice	✗	My favorite color is blue	✗
Human A	I have never had that, but I want to try it now.	I am a great cook	✗	I am a doctor	✗
Human B	What do you like to do the most?	I do whatever it takes to get what I want	✗	I am studying to be a dentist	✗

Figure 1: Black-box persona inference attacks (over 4,332 personas) on a dialog. Every representation of the utterance, which is based on the last hidden state of GPT-2, is attacked without defense (column of "Attacks on LM") and with defense (column of "Attacks on the defensed LM"). If the model can predict the persona of the speaker based on the observed representation, then we regard it as a successful attack; otherwise, unsuccessful. In practice, when deploying a model, a robust model which will reveal nothing of the encoded utterances is expected.

speakers' private persona attributes. Thus, it is necessary to improve training algorithms to address such overlearning issues. Finally, we apply defense learning strategies on the GPT-2 to prevent such black-box attacks. We combine proposed KL divergence loss (KL loss) with mutual information loss (MI loss) (Song et al., 2019) as additional defense objectives to train the GPT-2 and decrease the attacker's persona inference accuracy to 0.53%. Our contributions can be summarized as follows:[1]

1): To the best of our knowledge, we are the first to disclose and analyze the persona inference attack for LM-based chatbots and treat it as a privacy risk.

2): We propose an effective defensive training algorithm to prevent dialog representations from leaking personas of the corresponding speakers by uniform distribution approximation and mutual information minimization.

3): We conduct extensive experiments to quantify both privacy and utility of proposed defense mechanisms. Besides solving the persona leakage issue, the proposed training algorithm has nearly no negative influence on utility.

2 Related Work

Language models trained on private data suffer privacy risks of revealing sensitive information. Previous researches mainly considered black-box attacks that assumed attackers only had access to inputs and outputs of language models. Carlini et al. (2021) performed black-box model inversion attack on GPT-2 through descriptive prompts with beam search. Lehman et al. (2021) examined BERT pretrained on Electronic Health Records via blank filling and model probing to recover Personal Health Information. Furthermore, given black-box access to a language model's pre-train and fine-tune stages, Zanella-Béguelin et al. (2020) showed that sensitive sequences of the fine-tuning dataset can be extracted. For the distributed client-server setup, Malekzadeh et al. (2021) considered the sensitive attribute leakage from the server side with honest-but-curious (HBC) classifiers.

What is worse, for an LM-based chatbot, its training conversations are prone to include more private attributes than other commonly-used corpora for language modeling like BooksCorpus (Zhu et al., 2015) and Wikipedia. Tigunova et al. (2019) proposed Hidden Attribute Model (HAM) to extract professions and genders of speakers from various dialog datasets. Wu et al. (2020) further applied Attribute Extractor to generate speakers' attribute triplets flexibly and suggested downstream tasks based on the triplets. Pan et al. (2020) exploited embeddings of language models to recover inputs' digits and keywords. Though the setup of this work is similar to ours, they merely consider simple cases of data recovery with given rules and suffer great utility degradation to obtain optimal defense performance. For our work, there is no fixed pattern or rule for the model input. Instead of finding key-

[1]Code is publicly available at https://github.com/HKUST-KnowComp/Persona_leakage_and_defense_in_GPT-2.

words or recovering digits, we aim to infer more complicated private attributes from such embeddings. Moreover, our proposed defenses have almost no influence on the utility.

3 Attacking on Language Models

In this section, we illustrate black-box persona inference attacks on GPT-2 and our defense strategies. In Section 3.1, we first give the problem formulation. Then we describe the attack in Section 3.2.

3.1 Problem Formulation

We assume that there is a GPT-2 based chatbot f pretrained on private conversations D. Only language modeling is used to train the chatbot:

$$L_f(u;\theta_f) = -\sum_{i=1}^{|u|} \log(\Pr(w_i|c, w_0, w_1, ..., w_{i-1})), \quad (1)$$

where f refers to the LM-based chatbot with given utterance $u = \{w_0, w_1, ..., w_{|u|-1}\}$ and previous context c. An adversary owns one external annotated dialog dataset $D_a = \{(\mathbf{U}_1, \mathbf{s}_1), (\mathbf{U}_2, \mathbf{s}_2), ..., (\mathbf{U}_n, \mathbf{s}_n)\}$ with n conversations where $\mathbf{U}_i$ indicates a list of utterances $\{u_{i1}, u_{i2}, ..., u_{in_i}\}$ of i-th conversation and $\mathbf{s}_i$ corresponds to a list of sensitive personas $\{s_{i1}, s_{i2}, ..., s_{in_i}\}$ for corresponding utterance. Each persona s_{kj} is an integer that can be mapped to its persona according to a predefined dictionary and $0 \leq s_{kj} \leq C-1$ where C is the total number of predefined persona attributes. The goal of the adversary is to infer speakers' personas s from utterances' embeddings $f(u)$ where u and s refer to any utterance and its persona label.

3.2 Black-box Persona Inference Attack

The persona inference attack can be viewed as a supervised classification task. For the black-box attack setup, the adversary can only query the target dialog model f with access to embeddings of adversary's inputs and cannot access or modify model parameters θ_f. As shown in the left part of Figure 2, the adversary tries to build its attacker model $\mathcal{A}$ with its external data D_a and dialog model f. The persona predictor's output $\mathcal{A}(f(u))$ is the estimated probability distribution over C persona attributes. Its loss function $L_\mathcal{A}$ exploits cross-entropy between the predicted distribution and ground truth distribution that can be formulated as:

$$L_\mathcal{A}(u_{kj}, s_{kj}; \theta_\mathcal{A}) = \mathrm{CE}(\mathcal{A}(f(u_{kj})), s_{kj}), \quad (2)$$

where CE refers to cross-entropy loss between persona label s_{kj} and $\mathcal{A}(f(u_{kj}))$.

A well-performed persona predictor $\mathcal{A}$ can cause great privacy threats. For machine learning as a service (MLaaS), $\mathcal{A}$ can be applied to perform a man-in-the-middle attack on the application programming interfaces. Moreover, even if the raw data are protected and the transmission channel is secure, a curious service provider can train its attacker $\mathcal{A}$ to collect personas of service users.

4 Defense Learning Strategies

The LM training objective in Equation 1 only considers the utility of chatbots. In later experiment sections, we show that LM brings severe overlearning issues. Ideally, to achieve an optimal privacy-preserving chatbot against persona inference attacks, the probability distribution of the attacker model $\mathcal{A}$ should be close to the uniform distribution. That is, the adversary cannot improve its inference accuracy from posterior estimation $\mathcal{A}(f(u))$ and the accuracy is no better than making random guesses on the persona attributes. Moreover, the constraints on privacy should have minor degradation on the utility to maintain the strong generation ability of chatbots.

Following the intuition that the adversary cannot obtain better results than a random guess, in Section 4.1, we propose KL loss that aims to flatten the persona predictor's estimated distribution. Based on minimizing the mutual information between hidden states $f(u)$ of chatbots and private persona attributes s, we propose MI loss in Section 4.2. Lastly, we show the overall training objective in Section 4.3.

4.1 KL Loss

KL loss aims to minimize the Kullback–Leibler divergence between $\mathcal{A}(f(u))$ and the uniform distribution. It flattens the distribution of $\mathcal{A}(f(u))$ so that the adversary cannot gain any useful knowledge after training attacker model $\mathcal{A}$. The KL divergence between the uniform distribution and $\mathcal{A}(f(u))$ can be formulated as:

$$D_{KL}(\mathrm{UNI}||\mathcal{A}(f(u))) = -\frac{1}{C}\sum_{k=0}^{C-1} \log(C\Pr(k|f(u), \theta_\mathcal{A})), \quad (3)$$

where UNI indicates the uniform distribution and k indicates the k-th persona label of C labels. For optimization, we can leave out constant terms and the logarithm (Mireshghallah et al., 2021) to obtain

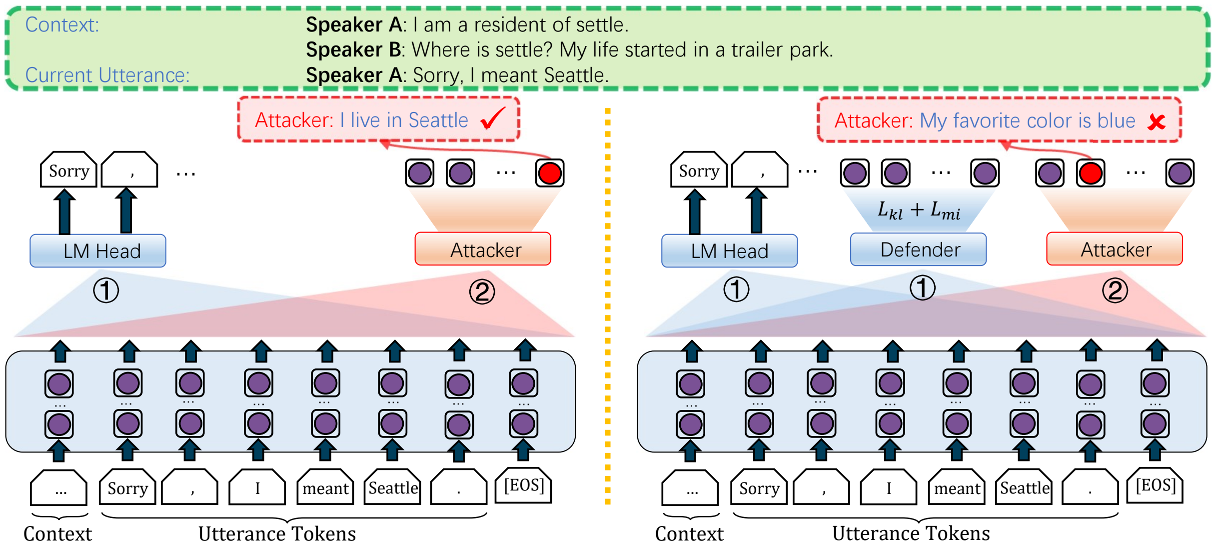

Figure 2: Scenarios for attacks without defense (left) and with defense (right). The GPT-2's training stage is marked by ① and the attacking stage is marked by ②. Both language modeling and defender objectives are jointly trained for the defense to optimize the GPT-2 model. After GPT-2's training stage ① is finished, parameters of GPT-2 are all frozen and then the attacking stage ② starts. The defender shares the same architecture as the attacker and uses L_{kl} with L_{mi} as defense objectives.

the following loss function:

$$L_D(u;\theta_{\mathcal{A}}) = -\frac{1}{C}\sum_{k=0}^{C-1} \Pr(k|f(u),\theta_{\mathcal{A}}). \tag{4}$$

However, from the perspective of defenders, they have no access to attacker model $\mathcal{A}$ and its parameters. Instead, they can build their own persona predictor as a fake attacker. More specifically, they may mimic the adversary to annotate a dataset D'_a and a persona predictor $\mathcal{A}_p$. Then the KL loss becomes:

$$L_{kl}(u;\theta_{\mathcal{A}_p},\theta_f) = -\frac{1}{C}\sum_{k=0}^{C-1} \Pr(k|f(u),\theta_{\mathcal{A}_p}), \tag{5}$$

where parameters of the chatbot θ_f and the fake attacker $\theta_{\mathcal{A}_p}$ are updated via KL loss. The intuition is to train the chatbot together with a fake attacker to prevent model overlearning by flattening the attacker model's distribution.

4.2 MI Loss

The privacy constraint requires that hidden representations should not reveal the persona attributes. In other words, given any utterance u and persona s behind the utterance u, we want to minimize the mutual information between $f(u)$ and s:

$$\min_{\theta_f} I(f(u);s). \tag{6}$$

Following the derivation in Song et al. (2019) and Li et al. (2020), the upper bound can be formulated as:

$$I(f(u);s) \leq \mathbb{E}_{q(f(u))} D_{KL}(q(s|f(u))||p(s)), \tag{7}$$

where $p(s)$ can be any distribution for s, $q(x)$ refers to probability distribution of model f parameterized by θ_f and $f(u)$ is assumed to be sampled from the conditional distribution $q(f(u)|x,s)$. However, $q(s|f(u))$ is hard to estimate. Instead, we use $p_\Psi(s|f(u))$ to approximate $q(s|f(u))$ via minimizing their KL divergence and then we can obtain the following lower bound (Song et al., 2019):

$$\begin{aligned}&\mathbb{E}_{q(f(u))} D_{KL}(q(s|f(u))||p(s))\\ &\quad\geq \mathbb{E}_{q(f(u))}[\log p_\Psi(s|f(u)) - \log p(s)].\end{aligned} \tag{8}$$

Therefore, our objective in Equation 6 can be formulated as an adversarial training objective:

$$\min_{\theta_f}\max_{\Psi} \mathbb{E}_{q(f(u))}[\log p_\Psi(s|f(u)) - \log p(s)]. \tag{9}$$

$\log p(s)$ is independent of $f(u)$, and we may leave this term out in Equation 9:

$$\min_{\theta_f}\max_{\Psi} \mathbb{E}_{q(f(u))}[\log p_\Psi(s|f(u))]. \tag{10}$$

Then, Equation 10 illustrates an adversarial game between an adversary p_Ψ who manages to infer s from $f(u)$ and a defender who modifies θ_f to protect s from persona inference attack. Adversarial training is widely used to protect sensitive features in natural language processing (Elazar and Goldberg, 2018; Coavoux et al., 2018; Li et al., 2018). Using the persona predictor model $\mathcal{A}_p$ with softmax activation to learn p_Ψ, we obtain the final objective for the defender:

$$\min_{\theta_{\mathcal{A}_p}} \max_{\theta_f} \mathrm{CE}(\mathcal{A}_p(f(u)), s). \tag{11}$$

We can rewrite Equation 11 into two losses: $L_{mi1}(u_{kj}, s_{kj}; \theta_{\mathcal{A}_p}) = \mathrm{CE}(\mathcal{A}_p(f(u_{kj})), s_{kj})$ and $L_{mi2}(u_{kj}, s_{kj}; \theta_f) = -\mathrm{CE}(\mathcal{A}_p(f(u_{kj})), s_{kj})$ for the fake adversary and the chatbot respectively. Then our MI loss can be formulated as:

$$L_{mi} = \lambda_0 L_{mi1} + L_{mi2}, \tag{12}$$

where λ_0 controls the ratio between two the fake attacker $\mathcal{A}_p$ and the defensed chatbot f.

4.3 Overall Objective

The right part of Figure 2 illustrates how the chatbot is trained to address the black-box attack. The loss function for the defender combines KL loss, MI loss and LM loss. Notice that the fake adversary objective in MI loss violates KL loss which tries to make the distribution of $\mathcal{A}_p$ flatten. Our proposed loss assigns more weights to the KL loss:

$$L = L_f + \lambda_1 L_{kl} + \lambda_2 L_{mi}, \tag{13}$$

where λ_1 and λ_2 are hyper-parameters and $\lambda_1 \geq 10\lambda_2$ to flatten the distribution of $\mathcal{A}_p$. Though the chatbot trained with overall loss L still cannot interfere training process of $\mathcal{A}$ during black-box attacks, L aims to mitigate persona overlearning issues of f to address such persona inference attacks.

5 Experiments

In this section, we conduct experiments to evaluate the performance of privacy and utility for the proposed defense learning strategies. In Section 5.1, we give our experimental settings in detail. In Section 5.2, we show the attacking performance with and without defense. In Section 5.3, we perform ablation study on defense objectives. In Section 5.4, we use automatic metrics to evaluate chatbots' utility. We conduct various attack setups in Section 5.5 and perform a case study in Section 5.6.

5.1 Experimental Settings

Dataset. To train the GPT-2 as our chatbot, we use the DialoGPT (Zhang et al., 2020) pretrained on Reddit comment chains. Then we use PersonaChat dataset (Zhang et al., 2018) to fine-tune the GPT-2. To obtain annotated dataset D_a for the adversary, we align personas to corresponding utterances through positive (utterance,persona) pairs provided in Dialogue NLI (Welleck et al., 2019) dataset. For those utterances with no annotations, we assign label -1 to them. We reshuffle the dataset to balance the label distribution among train/val/test datasets with the ratio of $8:1:1$. We first let the attacker and defender share the same training data. In later sections, we will separate the annotated data for the adversary and defender with no overlap. A summary statistics of D_a is shown in Table 1.

Stat Type	Value
Dialogs	10,907
Utterances (turns)	162,064
Unique personas	4,332
Total personas	98,056
Labeled turns	32,147
Avg. turns per dialog	14.86
Avg. labeled turns per dialog	2.95
Avg. words per turn	11.71

Table 1: Statistics of the aligned dataset.

Attacker model. In our experiment, we use a 2-layer neural network with cross-entropy loss as the attacker model. The attacker model exploits the final layer embedding of the last token "<|endoftext|>" from the GPT-2 as model input. We also try other attacker model architectures (transformer block based attackers) and input embeddings (average of all embeddings in the final layer of GPT-2), but the attacking performance is worse than the 2-layer model mentioned above.

Evaluation Metrics. The evaluation metrics are based on privacy and utility. For privacy, we use persona inference accuracy and weighted F1-score to evaluate the attacker's performance. We also use Bayesian Privacy (BP) (Gu et al., 2021a) to quantify the attacker's privacy loss for the estimated persona distribution. Top-k accuracy is reported in the Appendix. For utility, we apply BERTScore (Zhang* et al., 2020), Distinct (Li et al., 2016), BLEU (Papineni et al., 2002) and perplexity (PPL) as evaluation metrics. BERTScore and BLEU measure similarity between generated outputs and ground truth while Distinct (Dist) focuses on diversity. Perplexity shows the uncertainty when the LM model fits the data.

5.2 Privacy

Attacks without Defense. We list the attacking performance of $\mathcal{A}$ in multiple scenarios shown in Table 2. To demonstrate the overlearning issue of GPT-2, we consider 2 baseline attacks. If the adversary has no knowledge about persona attributes

	Acc	F1	Max-Ratio
Random Pred	0	0	0.02
Best Guess	0.72	1.02e-3	100
LM	37.59	3.65e-1	1.34
LM+KL+MI	0.53	6.78e-5	81.87
LM+KL	14.43	1.13e-1	10.60
LM+MI	0.53	5.57e-5	99.84

Table 2: Evaluation on the privacy over 4,332 persona labels. *Acc* and *Max-Ratio* are measured in %. *Acc* refers to test persona inference accuracy. *F1* uses weighted average F1-score. *Max-Ratio* indicates the ratio that the most frequent prediction shares among all predictions. The worse the attack model performs, the better privacy protection can be achieved.

distribution, then it can randomly guess over 4,332 labels (*Random Pred*). Otherwise the adversary can perform *Best Guess* that only guesses the most frequent persona in the dataset. *LM* indicates the attacker performance that only language modeling objective is applied to train the chatbot without any defense mechanism. From the table, the test persona inference accuracy on the *LM* achieves 37.59% while guessing on the label with most occurrences merely has 0.72% accuracy. That is, the black-box persona inference attack has $52\times$ the accuracy of guessing. The huge performance gap between the attacker model and the baseline guess method indicates that simple language modeling objective has serious overlearning issues that unintentionally capture private personas of speakers.

Attacks on the Defensed LM. To avoid the persona overlearning issue, we use additional defense objectives illustrated in Section 4. *LM+KL+MI* utilizes language modeling, KL loss and MI loss in Equation 13 to train the GPT-2. As demonstrated in Table 2, the attacker performance on *LM+KL+MI* significantly reduces the attacking accuracy from 37.59% to 0.53% and F1-score drops from 0.37 to nearly 0. This defense mechanism can even outperform *Best Guess* in terms of privacy protection. That is, even if the adversary annotates its own dataset to train an attacker model, the attacking performance is still worse than simply guessing the most frequent label. As a result, the black-box persona prediction attack becomes useless after applying the defenses for the chatbot. The adversary cannot obtain any speaker's persona from the embedding $f(u)$ by training $\mathcal{A}$.

To learn why the proposed defenses work so well, we further examine the ratio of the most frequent predicted label (*Max-Ratio*) among all predictions. The accuracy of *Best Guess* reveals that the most frequent label in the test set has a ratio of 0.72%. After applying KL loss and MI loss, the attacker model tends to make predictions on a single label. For *LM+KL+MI*, the *Max-Ratio* even occupies 81.87% predictions. This implies that the proposed defense strategies may have the potential to fool the attacker model to make wrong predictions on a single slot. We will further investigate this implication in later sections.

Overall, the above experiment demonstrates that our proposed defense learning strategies can effectively mitigate the persona overlearning issue and avoid black-box persona inference attacks.

5.3 Ablation Study

To show the effectiveness of proposed KL loss and MI loss and how they affect the performance of black-box persona inference attacks, we consider the inclusion and exclusion of proposed defense objectives. The result is shown in Table 2. *LM+KL* indicates the GPT-2 is trained with language modeling and KL loss. *LM+MI* applies language modeling and MI loss. From the table, it can be seen that *LM+KL*, *LM+MI* and *LM+KL+MI* are all able to reduce the test accuracy of the attacks. The KL loss is weaker from the perspective of defense, but it tends to flatten the estimated persona distribution with much smaller *Max-Ratio*. The *LM+MI* shares similar test accuracy and F1-score with *LM+KL+MI*, but nearly all predictions are made on a single persona label with a ratio of 99.84%. This suggests that MI loss causes the attacker model to predict all labels on a single persona attribute. After KL loss is applied on *LM+KL+MI*, the *Max-Ratio* drops to 81.87%.

As discussed earlier, high *Max-Ratio* may also cause privacy leakage. Suppose the adversary knows the persona with *Max-Ratio*, then it can improve its guess by not predicting this persona, which is a threat for fewer persona labels (for example, binary classification). These results verify that KL loss introduces flatter estimation and MI loss is more effective against persona overlearning, which conforms to our intuition of their objectives in Section 4.

5.4 Utility

Besides privacy, utility is another key objective to train a chatbot. Several automatic metrics are considered to evaluate the generation performance. For generation, we use GPT-2 to generate responses of

	PPL	Distinct		BLEU			BERTScore		
		Dist-1	Dist-2	BLEU-1	BLEU-2	BLEU-4	Precision	Recall	F1
LM	14.821	0.952	0.879	0.121	0.0551	0.0123	0.860	0.843	0.851
LM+KL	28.926	0.954	0.880	0.121	0.0558	0.0130	0.859	0.844	0.851
LM+MI	18.740	0.953	0.880	0.118	0.0531	0.0121	0.859	0.843	0.851
LM+KL+MI	19.674	0.953	0.880	0.119	0.0525	0.0105	0.858	0.842	0.850

Table 3: Evaluation on the utility over 4,332 persona labels.

	Unseen (0-2)				Overall (0-7)			
	Acc	F1	Max-Ratio	BP_u	Acc	F1	Max-Ratio	BP_u
Random Pred	34.42	0.35	33.90	0	13.21	0.14	13.35	0
Best Guess	56.84	0.41	100	2.60e-1	22.67	0.08	100	2.60e-1
LM	86.83	0.91	50.72	2.11e-3	72.37	0.72	20.94	3.04e-3
LM+KL+MI	28.68	0.37	58.15	2.84e-4	30.26	0.21	77.94	2.65e-4

Table 4: Evaluation on the privacy for 8 clusters. *Unseen* shows the results only for the first 3 persona labels that defender has never seen. *Overall* refers to the results on all 8 labels. *Acc* and *Max-Ratio* are measured in %. BP_u corresponds to Bayesian Privacy loss on the uniform distribution. Still, the worse the attack model performs, the better privacy protection can be achieved.

the second speaker (*Human B* in Figure 1) with all previous turns as context. Then we compared the generated model outputs with ground truth replies. We use *Dist-1* and *Dist-2* to count ratios of distinct unigrams and bigrams. *BLEU-1*, *BLEU-2* and *BLEU-4* are applied to evaluate generation similarity with ground truth. Due to the one-to-many nature of chit-chats, the *BLEU* is not adequate to compare generated responses with ground truth. Hence, we adapt *Precision*, *Recall* and *Precision* of *BERTScore* to measure the similarity in the embedding space.

The evaluation result is shown in Table 3, where same models from Table 2 are evaluated. The result indicates that adding KL loss will increase the perplexity greatly from 14.8 to 28.9. After combining KL loss with MI loss, its perplexity decreases to 19.674. A plausible explanation is that KL loss confuses the persona predictor and indirectly increases the uncertainty of the GPT-2. All GPT-2 models have relatively low *BLEU* scores due to the one-to-many mapping between contexts and responses. For *Distinct* and *BERTScore*, there are only minor differences between LM and defensed LMs. Though the uncertainty increases after applying KL loss and MI loss, it does no harm to the quality of generation. In summary, there is almost no negative influence on the utility after applying the proposed defense strategies.

5.5 More Setups on Attacks

Attacks on Imbalanced Data Distribution. Previous black-box attacks usually assume that the annotated dataset D_a must share similar data distribution with the defender's training data. To examine the performance of defense strategies on unseen personas, we assign the adversary's dataset D_a with labels that the defender cannot acquire. We split data with 500 persona labels that are uniquely held by the adversary. The defender owns 8,031 conversations with persona labels ranging from 500 to 4,331 while the adversary holds 2,376 dialogues with persona labels ranging from 0 to 4,331. For testing, 500 conversations with persona labels ranging from 0 to 4,331 are used.

Under imbalanced data distribution, the attack on the defensed LM has *Acc* 0.47%, *F1* 1.90e-3 and *Max-Ratio* 94.06%. The persona inference accuracy is still very low and the attacker model tends to predict more on a single persona label than the balanced data distribution setup. This result shows that the proposed overall loss can also prevent black-box persona inference attacks on unseen personas. It also verifies previous suggestions that combining LM loss with MI loss may fool the attacker model to make wrong predictions.

Attacks on Fewer Persona Labels. The above experiments are based on 4,332 persona labels. In fact, many personas share similar meanings and can be further clustered. Besides, to better evaluate privacy loss for the estimated distribution, a smaller label space is preferred. Therefore, it is necessary to consider defense performance on a smaller label space. We use Sentence-BERT (Reimers and Gurevych, 2020) to embed all persona sentences and perform k-means clustering on the embeddings to obtain 8 clusters. We manually checked these clusters and classified them as cars, food, animals

Context: 1): **Bot**: Hi, how are you doing? 2): **User**: Not great! I hate school.
3): **Bot**: Oh no, what's wrong with your school? 4): **User**: I just do not like it! Plus I am so worried about money.

	Generation & attack on LM				Generation & attack on the defensed LM		
5): **Bot**	Oh that is not good, I am in the military.	I am in the army	✓	5): **Bot**	I understand that, I am too.	I love the beach	✗
6): **User**	What do you do in the military?	I am in the army	✗	6): **User**	What do you like to do?	I love animals	✗
7): **Bot**	I am in the navy, the food is so amazing.	I would be honored to give my life for my country	✗	7): **Bot**	I love to bake and spend time with my family.	My favorite color is blue	✗
8): **User**	Do you have any children?	I want to have two kids	✗	8): **User**	I like to watch game of thrones.	My favorite color is blue	✗
9): **Bot**	Yes, they are all grown up.	I have a son	✗	9): **Bot**	I have never seen it before.	I have my own salon	✗
10): **User**	Well, I like to watch game of thrones with my family.	I love watching game of thrones	✓	10): **User**	Do you have any favorite music?	My favorite color is blue	✗

Figure 3: Black-box persona inference attacks on chit-chats between the user (the authors of this paper) and chatbots. For both conversations, the "context" is fixed and used as the first four utterances. Then the bot and the user start interactive conversations with the "context". Since there is no gold standard, the results are annotated by the authors.

(pets), family information, hobbies, jobs, personal information and music tastes respectively. To evaluate how the clustering performs, we randomly sample 100 utterances with clustered labels and invite two volunteers to inspect those samples. Both of them agree on 90% of the clustered annotations. After manual inspection of the remaining 10% annotations, the clustering error rate is 8%. Following previous imbalanced data split, we assign data in the first 3 clusters only to the adversary to make the data distribution imbalanced. Here, the defender owns 6,654 conversations with persona labels ranging from 3 to 7 while the adversary holds 3,753 dialogues with persona labels ranging from 0 to 7. For testing, 500 conversations with persona labels ranging from 0 to 7 are used.

The attacking performance for both unseen labels and all labels is displayed in Table 4. BP_u measures the KL divergence $D_{KL}(F_0||\mathcal{A}(f(u)))$ where F_0 refers to uniform distribution. For imbalanced data distribution with a small label space, our proposed defenses can still achieve much lower attack accuracy than *LM* on both *Unseen* and *Overall*. However, for *Overall*, *LM+KL+MI* has higher accuracy with a lower F1-score compared with two baselines. This indicates that proposed defenses fail to protect privacy as we desired in the baselines. For BP_u, *LM+KL+MI* are around 10 times smaller than *LM*. It means that after applying defense objectives, the attacker's estimated distribution is much closer to the uniform distribution. Thus the effectiveness of the KL loss is verified. In addition, *Max-Ratio* with 8 clusters on *Unseen* is smaller than 4,332 labels even though the distribution of 8 clusters is obviously tighter. Still, the *Max-Ratio* of 58.15% accounts for a much larger fraction than other predictions. In summary, the above results imply that for the smaller label space, our proposed defense objectives are still effective even on unseen persona labels.

5.6 Case Study

In Figure 3, we give an example of the persona inference attack, where conversations are generated between the chatbot and the user with the given context. We manually mark True/False on the predicted results. As shown in the figure, there are several successful attacks on LM and no correct prediction on the defensed LM. For attacks on LM, speakers' hobbies and jobs can be inferred. For incorrect predictions, the attacker model can still predict context-aware personas. After applying proposed defense learning strategies, the predicted personas become irrelevant with context and mostly predict "My favorite color is blue." In fact, it is the most frequent prediction for *LM+KL+MI* over 4,332 persona labels. This attack example illustrates that our defense objectives can prevent the black-box persona inference attack from inferring relevant personas.

6 Conclusion

In this paper, we show that LM-based chatbots tend to reveal personas of speakers and propose effective defense objectives to prevent GPT-2 from overlearning. Unlike other works that suffer from utility degradation, our defense learning strategies do no harm to the powerful generation ability of LM-based chatbots. We conduct extensive experiments to evaluate both privacy and utility. We perform black-box persona inference attacks under various setups to demonstrate the robustness of proposed defense learning strategies. In addition, we use automatic metrics to show that proposed defense learning strategies maintain the utility. For future work, we suggest working on flattening the distributions of attacker models.

7 Ethical Considerations

We declare that all authors of this paper acknowledge the *ACM Code of Ethics* and honor the code of conduct. This work essentially considers black-box attacks on the private persona attributes and proposes effective learning strategies to prevent chatbots from overlearning private personas.

Dataset. During our dataset collection, all the conversations and personas are collected from publicly available datasets including PersonaChat and DNLI. All the speakers are anonymized and no identifiable personal information is included.

Model. For training our LM-based chatbots, we follow standard methods. We are well aware of the bias issue inside current language models. In the future, if there are other fairer language models, we will extend our defenses on them.

Acknowledgment

The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20) and the GRF (16211520) from RGC of Hong Kong, the MHKJFS (MHP/001/19) from ITC of Hong Kong and the National Key R&D Program of China (2019YFE0198200) with special thanks to Hong Kong Mediation and Arbitration Centre (HKMAAC) and California University, School of Business Law & Technology (CUSBLT), and the Jiangsu Province Science and Technology Collaboration Fund (BZ2021065).

References

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *Proceedings of USENIX Security Symposium*, pages 2633–2650.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018. Privacy-preserving neural representations of text. In *Proceedings of EMNLP 2018*, pages 1–10.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of EMNLP 2018*, pages 11–21.

Hanlin Gu, Lixin Fan, Bowen Li, Yan Kang, Yuan Yao, and Qiang Yang. 2021a. Federated deep learning with bayesian privacy. *arXiv preprint arXiv:2109.13012*.

Jing Gu, Qingyang Wu, Chongruo Wu, Weiyan Shi, and Zhou Yu. 2021b. PRAL: A tailored pre-training model for task-oriented dialog generation. In *Proceedings of ACL 2021*, pages 305–313.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of ACL 2020*, pages 583–592.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of ICLR 2020*.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of NAACL 2021*, pages 946–959.

Ang Li, Yixiao Duan, Huanrui Yang, Yiran Chen, and Jianlei Yang. 2020. Tiprdc: Task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations. *In Proceedings of the ACM SIGKDD 2020*, page 824–832.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL 2016*, pages 110–119.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of ACL 2018 (Volume 2: Short Papers)*, pages 25–30.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR 2019*.

Mohammad Malekzadeh, Anastasia Borovykh, and Deniz Gündüz. 2021. Honest-but-curious nets: Sensitive attributes of private inputs can be secretly coded into the classifiers' outputs. In *Proceedings of the ACM CCS 2021*.

Fatemehsadat Mireshghallah, Huseyin Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. 2021. Privacy regularization: Joint privacy-utility optimization in LanguageModels. In *Proceedings of the NAACL 2021*, pages 3799–3807.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *Proceedings of 2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of EMNLP 2020*, pages 4512–4525.

Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. OTTers: One-turn topic transitions for open-domain dialogue. In *Proceedings of ACL 2021*, pages 2492–2504.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI 2021*, volume 35, pages 13789–13797.

Congzheng Song and Vitaly Shmatikov. 2020. Overlearning reveals sensitive attributes. In *Proceedings of ICLR 2020*.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning controllable fair representations. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 2164–2173.

Anna Tigunova, Andrew Yates, Paramita Mirza, and G. Weikum. 2019. Listening between the lines: Learning personal attributes from conversations. In *Proceedings of WWW 2019*, page 1818–1828.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of ACL 2019*, pages 3731–3741.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint*, arXiv:1901.08149.

Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Getting to know you: User attribute extraction from dialogues. In *Proceedings of LREC 2020*, pages 581–589.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of NIPS 2019*, pages 5754–5764.

Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Ruehle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. 2020. Analyzing information leakage of updates to natural language models. In *Proceedings of ACM CCS 2020*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of ACL 2018*, pages 2204–2213.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR 2020*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of ACL 2020: System Demonstrations*, pages 270–278.

Yukun Zhu, Ryan Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of ICCV 2015*, pages 19–27.

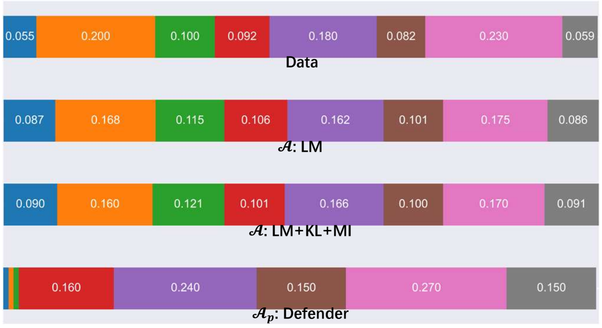

Figure 4: The test set data distribution (*Data*) and average estimated distributions of persona predictors ($\mathcal{A}$ and $\mathcal{A}_p$) over imbalanced 8 clusters.

A Training details.

For each conversation, the utterances are concatenated by the special token “<|endoftext|>” to train the GPT-2. To decode outputs from GPT-2, we apply the Nucleus Sampling (Holtzman et al., 2020) method. We set top-p = 0.9 with a temperature coefficient 0.9 to sample words from the GPT-2. For optimization, we set 2 AdamW optimizers (Loshchilov and Hutter, 2019) for the chatbot and the persona predictor respectively. The learning rate is 3e-5 with linear warm-up and decay. For hyper-parameters, we set $\lambda_0 = 1$, $\lambda_1 = 10$ and $\lambda_2 = 1$.

B Comparison of Internal Distribution between $\mathcal{A}$ and $\mathcal{A}_p$

To make predictions on personas, the $\arg\max$ function is used for the estimated distribution of persona predictors. However, the internal distribution conveys crucial information about how the persona predictors estimate $f(u)$. We follow the setup of imbalanced data split of 8 clusters in Section 5.5 to examine persona predictors of attacker $\mathcal{A}$ and fake attacker $\mathcal{A}_p$.

Figure 4 shows the data distribution of the test set and average distribution after softmax activation over the 8 labels for attacker $\mathcal{A}$ and defender $\mathcal{A}_p$. For attacker $\mathcal{A}$, we consider the attack on *LM* and *LM+KL+MI*. The defender $\mathcal{A}_p$ tends to have a large difference with *Data* and tries to flatten its distribution among its own training set (the last 5 labels). This behavior conforms to the KL loss's objective that aims to flatten the distribution and deviate from the ground truth distribution. For attacker $\mathcal{A}$, distributions of both *LM* and *LM+KL+MI* seem close to the ground truth distribution. This indicates that the attacker model $\mathcal{A}$ can still learn statistical information about personas. However, its attacking performance is poor. The poor performance implies our proposed defense learning strategies may obfuscate *Attacker* for estimating single sample $f(u)$ and finally make the wrong prediction.

Human A	Hello, how are you tonight?
Human B	Hello my friend. I am well.
LM	I am fine and you?
LM+KL+MI	Hi, I am good, you? What do you like to do for fun?
Human A	Good, glad to hear it. What do you do for fun?
Human B	I ride around the town on my cool bicycle.
LM	I like to go to the beach.
LM+KL+MI	I collect civil war antiques, and to read, you?
Human A	Really? I really like mountain bike too.
Human B	I wish I lived in the mountains.
Human A	Do you like nature? I have been to 12 national parks.
Human B	I love nature. I like looking at plants.
Human A	I love plants too, and hiking. In fact, I am actually an environmental activist.
Human B	Cool, I am a vegan.
Human A	Nice, do you have a favorite food?
Human B	My favorite dish is lentil curry.
LM	I do not like meat.
LM+KL+MI	I do not eat meat.
Human A	I have never had that, but I want to try it now.
Human B	What do you like to do the most?
LM	It is very good. I also really enjoy watching funny videos of comedians.
LM+KL+MI	It is amazing. My wife makes it.

Figure 5: Dialog generation example on *Human B*. All previous utterances between A and B are used as context to generate responses.

C More on Case Study

C.1 Example of Generation

To show an intuition view on utility, we provide one generation sample shown in Figure 5. Both *LM* and *LM+KL+MI* are able to generate fluent and proper relies. Moreover, they tend to maintain coherence with previous contexts. For example, it is mentioned in the context that *Human B* is a vegan and both chatbots respond that they do not eat meat for the food preference. This generation example shows that proposed defense learning objectives preserve the model utility.

C.2 More Examples of Persona Inference

Here, we give two more examples of the persona inference attacks in Table 6. The first example shows one successful defense. For the second example, both attackers with and without defense fail to predict the ground truth persona. Still, we can see that *LM+KL+MI* predicts personas that are irrelevant

	Top-1	Top-3	Top-5	Top-10	Top-50	Top-100	Top-500	Top-2000
LM	37.59	55.57	63.28	72.76	87.19	91.54	97.79	99.60
LM+KL+MI	0.53	1.80	2.24	3.20	8.64	12.10	30.57	80.22

Table 5: Evaluation on the top-k accuracy over 4,332 clusters.

Dialog Context 1 **Human A**: Hi how are you doing? **Human B**: I am great and you? **Human A**: I am great, just reading. **Human B**: I am listening to the rolling stones, I love them. **Human A**: Is that your favorite band? **Human B**: Yes it is. I am working right now too. **Human A**: Where do you work at? **Human B**: IBM in Chicago, what about you?
Persona Prediction: **LM**: I currently work for IBM in Chicago. ✓ **LM+KL+MI**: I love cats. ✗
Dialog Context 2 **Human A**: Hello there my name is Dr.Lucy. How are you? **Human B**: I am great, loving this city life, how are you? **Human A**: I am well thank you. I miss my country life in Spain. **Human B**: My older brother lives in Spain, how is it? **Human A**: It is beautiful. I hope to take my family back there. **Human B**: Yes, maybe i will take my girlfriend that I love there one day. **Human A**: Oh, how long have you two been together? **Human B**: Very long, she was with me when I colored my hair pink. **Human A**: That is awesome. What type of music do you two listen to? **Human B**: I like reading music, what about you? **Human A**: Hip hop is my favorite. Do you play an instrument?
Persona Prediction: **Ground truth**: My favorite music is hip hop. **LM**: I know how to play the guitar. ✗ **LM+KL+MI**: My favorite food is pizza. ✗

Table 6: More persona inference attack examples. The embeddings of the final utterance with orange color are used for inferring B's persona.

to the context. However, *LM*'s output "I know how to play the guitar." is much closer to the context about music and instruments. Without any defense, the above examples show that the attacker model can still predict context-aware personas even if its predictions are wrong. After applying the proposed defenses, the attacker model cannot predict meaningful personas relevant to the context.

D Evaluation on Top-k Accuracy

Previous experiments mainly consider accuracy as the evaluation metric. In this section, we use top-k accuracy for the black-box persona inference attacks to measure privacy protection. As shown in Table 5, our defense is much more robust than *LM* when k $\leq$ 50. When k is larger than 500, the defense degrades rapidly as k increases. This result implies that the ground truth personas mostly lie in the top 2,000 predictions even if the defense is applied. For a smaller k, our proposed defense learning strategies are still effective.