

# Detecting Unintended Memorization in Language-Model-Fused ASR

W. Ronny Huang, Steve Chien, Om Thakkar, Rajiv Mathews

Google LLC

{wrh, schien, omthkkr, mathews}@google.com

## Abstract

End-to-end (E2E) models are often being accompanied by language models (LMs) via shallow fusion for boosting their overall quality as well as recognition of rare words. At the same time, several prior works show that LMs are susceptible to unintentionally memorizing rare or unique sequences in the training data. In this work, we design a framework for detecting memorization of random textual sequences (which we call canaries) in the LM training data when one has only black-box (query) access to LM-fused speech recognizer, as opposed to direct access to the LM. On a production-grade Conformer RNN-T E2E model fused with a Transformer LM, we show that detecting memorization of singly-occurring canaries from the LM training data of 300M examples is possible. Motivated to protect privacy, we also show that such memorization gets significantly reduced by per-example gradient-clipped LM training without compromising overall quality.

**Index Terms:** LM Fusion, Unintended Memorization

## 1. Introduction

Neural networks for common ASR applications like smart assistants, dictation, and telephony, are trained on millions of utterances which can often contain sensitive information, such as credit card numbers. Given the prevalence of ASR in user products, ensuring the privacy of training data is an increasingly important topic in the speech community [1, 2, 3].

Modern E2E models dedicate most of their resources to encoding audio. For example, the Conformer(L) architecture [4] dedicates 17 layers to the encoder to process audio, and only 1 layer to the decoder to process text. This may appear to limit an E2E model’s ability to memorize its textual data. However, the ASR pipeline is often equipped with an accompanying language model (LM) whose job is to use its linguistic knowledge to disambiguate acoustically similar hypotheses given by the acoustic model (AM)<sup>1</sup>. As such, LMs are often integrated into production speech engines, improving WER by 5-10% relative [5] with larger gains on rare words [6].

For better domain match, such LMs are often trained on the transcripts of the same data as the AM. Additionally, personalized or contextually biased [7] ASR systems use a standard AM with an LM trained on per-user information (e.g. contacts list [8], typing patterns [9], etc.)—packing yet more personal information into the LM. Several prior works [10, 11, 12, 13, 14, 15] have shown that LMs are susceptible to unintentionally memorizing rare or unique sequences of data, thus making textual training data susceptible to such leakage. This leaves the open question, which we try to answer in this work: *Does adding*

*an LM to an AM create a privacy vulnerability in the combined ASR system even without direct access to the LM?*

Our first contribution is to design a framework to test for unintended memorization using only query access to the ASR system. We do this by inserting a set of random “canaries” into the LM’s training data, and observing the ASR model’s performance on these examples. We need to overcome a pair of obstacles: First, since ASR models accept audio and not text, we use a Text-To-Speech (TTS) system to generate synthetic utterances for querying the ASR system. Next, unlike conventional attacks on LMs which require some confidence score, e.g., perplexity, ASR systems typically only give the single top transcript for a given audio sequence. Thus, we use the model’s WER on canaries versus baseline examples as our primary metric.

We test for memorization across a range of canary frequencies and LM capacities. Our results show that memorization is detectable even for canaries that appear only once in a large (300M sample) training corpus, and then grows significantly the more frequently a canary is repeated in the training data (from 1 to 32 times). We also show how increasing model capacity leads to increased memorization.

Our next contribution is to study the effect of per-example gradient clipping on mitigating memorization, an idea motivated by prior work [10, 12]. Our experiments here demonstrate that this technique can significantly reduce the level of memorization without appreciable loss of overall utility in the model.

Finally, we observe that while using *clean* TTS utterances can be useful for detecting memorization, there can be instances where the ASR system may not leverage the LM for providing the top transcripts, e.g., if the TTS utterances are easy to predict for the AM itself. In such cases, the LM’s memorization may not be able to surface to the top transcript predictions which we use for inference. Thus, we add noise to parts of the TTS utterances such that the LM’s predictions can be leveraged by the ASR system. Moreover, we design a simple classifier that can use the top transcripts for these utterances to conduct a *membership inference* (MI) task. We provide an analysis of how metrics like precision and recall (commonly used for evaluating MI attacks [16, 17, 18]) can be used to demonstrate the extent of unintended memorization in ASR LMs. For instance, we show in our experiments that while our classifier shows a baseline (50% precision, 13% recall) for non-memorized sequences of data, it achieves (~80% precision, ~50% recall) for sequences that appeared only eight times among 300M other samples for LM training and were subsequently memorized.

**Related Work:** Recent work has shown that model updates during ASR training can leak potentially sensitive artifacts like labels [1] and speaker identity [2] of utterances used in computing the updates. There have been works [10, 11, 12, 13, 14, 15] that focus on designing extraction techniques to demonstrate the susceptibility of LMs to memorize training data. There are also works [19, 20] that theoretically study such memorization and how it relates to generalization in learning.

Submitted to Interspeech 2022.

<sup>1</sup>In this paper, we refer to an E2E model as an AM to better contrast it from an LM. We acknowledge the term “acoustic model” can historically refer to only the audio processing component, e.g., an encoder.

## 2. Method

### 2.1. Detecting unintended memorization

We start by defining a (publicly-known) distribution over the tokens of the model. Next, we draw some i.i.d. samples from the distribution and add these “canaries” with varying frequencies into the LM’s normal training set. Simultaneously, we also draw an equal-sized set of “extraneous” samples in the same manner, which are not used for training.

After training the LM, we query an AM+LM recognizer to detect the level of unintended memorization of the canaries. This is done by first using a WaveNet TTS system [21] to generate synthetic utterances for the canaries and extraneous examples. These are fed into the recognizer, and the top-1 transcripts are retrieved. We then use two methods to measure unintended memorization: First, we compare the word error rate (WER) of the canaries from a canary-trained model, and contrast it with the same measurement from an extraneous-trained model. Second, as described later in §2.3, we also design a membership classifier that modifies the TTS utterances to accurately predict memorization on a *per-example* basis.

### 2.2. Training using per-example gradient clipping

An established technique for reducing unintended memorization in ML models is DP-SGD [22, 23]. This algorithm modifies standard gradient descent in two ways: it first clips the gradient of each training example to a fixed maximum *clip norm*, and then adds random noise to the mean clipped gradient. A model trained this way is *differentially private* [24], which is a formal notion for quantifying privacy leakage. Unfortunately, large models trained using DP-SGD with strong privacy guarantees tend to have unacceptably low utility (i.e. generalization) compared with baseline (non-private) models.

A possible step forward is to only perform per-example clipping, thus bounding the *sensitivity* of each gradient. A model trained this way no longer satisfies any meaningful differential privacy, but prior work [10, 12] has shown that such models tend to match the utility of baseline models while being empirically less prone to memorization.

### 2.3. Membership classifier

A conventional membership inference attack uses a membership classifier that predicts whether or not a candidate example is in the training set based on the true label and the output posteriors of the model for that candidate example. In our case, there are two challenges. First, realistic speech recognizers only output the top-1 hypothesis; thus, this is the only artifact available to the membership classifier. This situation is akin to label-only membership inference attacks [25]. Second, the model we wish to probe here is the LM, but we can only send inputs to the AM. Thus, we must synthesize TTS audio of the candidate text sequence for which we wish to predict the membership. This presents its own set of challenges. On one hand, if the TTS is too clear, the AM alone might predict the sequence perfectly without the LM’s help. On the other hand, if the TTS is too unclear, the AM could output an empty transcript, inhibiting the LM’s ability to influence the prediction.

We opt for a middle ground: The first part of the candidate example, which we call the prefix, is synthesized with clear TTS, while the remainder, which we call the suffix, has Gaussian noise added. This partially obscured example is then fed into the fused model, and its output observed. From here, we use a very simple membership classifier: if the model’s output

exactly matches the ground truth, the classifier predicts that the example was used in training (i.e. is a canary); otherwise, the classifier predicts the example was not used in training.

## 3. Setup

### 3.1. Canaries

Each canary is a fixed-length sequence of random lower-case alphabetical characters with spaces between them, e.g. “o e g d b u”. This mimics the format of sensitive data like passwords, serial numbers, and other secret codes. Further, this format has a well-defined baseline likelihood of  $26^{-n}$ , where 26 is the number of possible characters, and  $n$  is the sequence length. This is useful for confirming the extent to which the LM has memorized the canary sequences by comparing their perplexity to  $26^n$ , which we do during training. For each canary, we insert it into the LM training data at a certain frequency (i.e., number of times it’s repeated). We choose frequencies that are powers of 2, as shown in Table 1. For example, there are 16384 canaries that appear once in the training set (CAN1), 8192 canaries that appear twice (CAN2), and so on until CAN32. We also keep a set of canaries that are not inserted into the training set (CAN0) to use as a baseline during evaluations.

Table 1: *Canary datasets. Each dataset is inserted into training at the frequency indicated.*

Name	CAN0	CAN1	CAN2	CAN4	CAN8	CAN16	CAN32
Frequency	0	1	2	4	8	16	32
# Canary	16384	16384	8192	4096	2048	1024	512
# Extraneous	16384	16384	8192	4096	2048	1024	512

We also craft another equally-sized set of *extraneous* sequences that take the same structure (e.g. 6 spaced random letters) as the canaries as well as the same frequency distribution. The only difference is that their random sequences are sampled independently and contain no matches with the canaries. The extraneous set is used for training a baseline LM that has a similar distribution as the LM trained with canaries. In addition, the extraneous set is used to compute the precision and recall metrics discussed in §4.4.

### 3.2. Dataset

Our main ASR task is voice search (VS); as such, our test set consists of 11k utterances from Google’s voice search traffic. The utterances are only a few words long. The training set for both the AM and LM is a slice of Google search traffic from multiple domains such as voice search, farfield, telephony, and YouTube, making up ~300M utterances. All utterances are anonymized and hand-transcribed, with the exception of YouTube being semi-supervised. Our data collection abides by Google AI Principles [26]. This is the same training set used in Google’s production speech [27] and language [28] models. For the speech model, the acoustics are further diversified via multi-style training [29], random down-sampling from 16 to 8 kHz, and SpecAug [30]. The LM uses only the transcripts of the training set. The canaries are merged into the LM’s training set only (the AM is not trained on canaries), and make up less than 0.1% of the total.

### 3.3. Architecture

Our acoustic model is the same as in [31]. It is a state-of-the-art Conformer-encoder, streaming RNN-T model with 150M parameters emitting 4096 lower-case wordpieces. The decoder is

a tiny 2M-parameter stateless embedding decoder [32] with no recurrence. Our language model is a 70M-parameter, 10-layer transformer [33] with model dimension of 768, feed-forward dimension of 3072, and 5 attention heads. It is trained using the AdaFactor optimizer with a batch size of 1024 for 800k steps.

### 3.4. Language model fusion

Shallow fusion [34], i.e. interpolation of the AM and LM logits during decoding time, is the predominant LM fusion method used in performance speech recognition and is the method we use here. In particular, we apply HAT shallow fusion which includes an additional interpolation term that factorizes out the AM’s log-posterior from its internal language model score [35]. In summary, the decoder optimizes

$$y^* = \underset{y}{\operatorname{argmax}} [\log p(y|x) - \lambda_2 \log p_{\text{ILM}}(y) + \lambda_1 p_{\text{LM}}(y)],$$

where  $p(y|x)$  is the AM log-posterior score,  $p_{\text{ILM}}(y)$  is the AM’s internal language model score, and  $p_{\text{LM}}$  is the LM score. The interpolation weights ( $\lambda_1, \lambda_2$ ) are optimized via Vizier [36] where the objective is to reduce WER on the VS test set.

## 4. Experiments

### 4.1. Baseline comparisons

Table 2 shows the performance of a collection of recognizers, each consisting of a different LM fused with the same AM. Since the AM is fixed for all experiments, we name each row by the LM for brevity; note that each row represents the decoded WER results from an entire AM+LM recognizer. For each recognizer, we decode on the VS test set to measure overall recognizer quality. We also decode on the TTS of the single-frequency canaries, CAN1, as well as CAN0 as a baseline. Recall §3.1 and Table 1 for discussion on these canary datasets.

Table 2: WER on VS, CAN0, and CAN1 for different LMs. Lower values indicate more generalization (VS) or memorization (CAN0 or CAN1). Naming convention for LMs is {size}-{trained\_on\_canary\_or\_extraneous\_set}.

LM	VS	CAN0	CAN1
B1: None	6.53	34.02	33.98
B2: 70M	6.22	33.88	34.50
B3: 70M-EXT	6.24	23.59	24.21
E1: 9M-CAN	6.51	20.93	21.20
E2: 30M-CAN	6.25	22.13	21.43
E3: 70M-CAN	6.26	22.97	21.17
E4: 130M-CAN	6.22	23.46	20.77

We first introduce several baselines. B1 is the AM alone without an accompanying LM, and B2 is the AM fused with a 70M parameter LM trained on the same data. B2 provides a 5% WER relative improvement on the VS test set over B1, confirming that LM shallow fusion is properly configured to provide overall gains. B3’s LM is also 70M parameters but trained on the extraneous (EXT) dataset; thus it has learned the same *structure* as the canaries (6 space-separated alphabetical letters), but not their specific *sequences*. As a result, B3’s CAN0 and CAN1 WERs are 10% lower than that of B1 and B2, which were not trained on the same structure. B3 is the most appropriate baseline for comparing our canary-trained LM results as it factors out the memorization of canary *sequence* from the learning of canary *structure*.

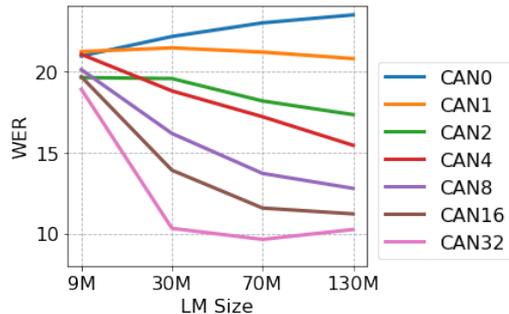


Figure 1: WER on canary sets of various frequencies as a function of LM size. Lower values indicate more memorization.

### 4.2. Unintended memorization

Experiments E1-E4 in Table 2 consist of different-sized LMs (9M, 30M, 70M, 130M) trained on the canaries. Each size was obtained simply by applying a constant multiplier to the number of attention heads, model dimension, and feed-forward-layer dimension of the Transformer architecture. As a sanity check, training on canaries does not significantly compromise overall performance as indicated by the fact that VS WER is comparable to B2 and B3. The exception is E1, which provides no gain to VS due to its small LM size.

On CAN0 (canaries not in training) the WER increases as the model size increases (E1-E4). This is explained by the fact that larger LMs have greater modeling power, and stronger LMs are better able to assign high perplexities to random sequences they have not seen in training. On the other hand, larger LMs are also better able to assign low perplexities to sequences they *have* seen in training. This pattern is evident for single-frequency canaries in the CAN1 column where WER decreases with larger models. In Figure 1, we plot the WER versus LM size dependence for all canary frequencies. As expected, the trend of increased memorization (lower WER) with LM size is even stronger for more frequent canaries (CAN2-CAN32). Thus, while increasing LM model size is a common way to improve performance, unfortunately this has the side-effect of making unintended memorization worse.

### 4.3. Mitigation via per-example gradient clipping

Now that we have shown that LMs can memorize even single frequency canaries in a large dataset, we focus on a strategy to mitigate the memorization. In Table 3, we report the results on different levels of per-example gradient clipping (discussed in §2.2). At a clip norm of 32, no gradients are clipped. Thus, the top row can be seen as a non-clipped baseline that is most prone to memorization, similar to the models in §4.2. At a clip norm of 0.5, 99% of the per-example gradients get clipped. At each clip norm level, we train an LM on the canaries (70M-CAN), along with a corresponding LM on the extraneous set (70M-EXT) as a baseline. This allows us to make well-baselined measurements of canary memorization, as the only difference between the WER given by 70M-CAN and 70M-EXT is due to the memorization of the canary *sequences*. All other factors, including the canary *structure* and clip norm level, are identical between 70M-CAN and 70M-EXT for each row.

A first observation is that the VS WER is mostly unchanged regardless of the clip norm level, indicating that gradient clipping does not compromise model quality. Next, on CAN1, 70M-CAN obtains lower WERs compared to the correspond-

Table 3: WER of VS and CAN1 at different levels of gradient clipping. Each testset is evaluated on 70M-CAN and 70M-EXT. WERR is WER relative of 70M-CAN vs. 70M-EXT. Less negative values of WERR indicate more memorization mitigation.

Clip norm	VS		CAN1		WERR (%)
	70M-EXT	70M-CAN	70M-EXT	70M-CAN	
32	6.24	6.26	24.21	20.45	-15.5
16	6.23	6.30	25.16	22.55	-10.4
8	6.27	6.22	25.12	23.10	-8.0
4	6.27	6.29	24.79	23.48	-5.3
2	6.26	6.29	25.00	23.39	-6.4
1	6.30	6.23	24.93	23.37	-6.3
0.5	6.19	6.28	24.39	23.50	-3.6

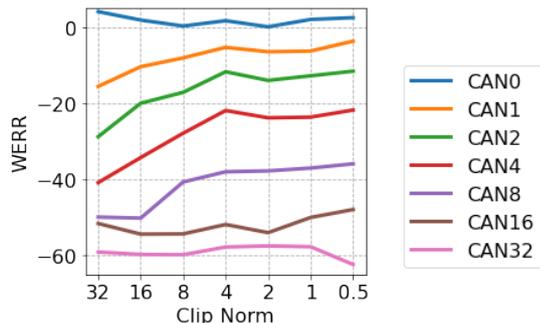


Figure 2: Percent WER relative (70M-CAN vs. 70M-EXT), or WERR, on canary sets as a function of the gradient clipping level. More negative WERR indicates more memorization. Conversely, less negative WERR indicates stronger mitigation.

ing 70M-EXT baseline at each clip level due to memorization. The amount of memorization can be measured as the WER relative of 70M-CAN with respect to 70M-EXT, and we display this value in the last column (WERR). WERR starts at -15.5% at a clip norm of 32 and reduces to -3.6% at a clip norm of 0.5, making clear that the models trained with more aggressive clipping (lower clip norm) can achieve significantly less memorization.

Figure 2 plots the WERR for canaries of all frequencies (CAN0-CAN32) with respect to clip norm. Up to CAN8 (frequency of 8), there is a trend of memorization decreasing (WERR becoming less negative) with more aggressive clipping (smaller clip norm). Beyond that (CAN16 and CAN32), canaries are so frequent that they might be considered data to be learned rather than idiosyncratic examples. We see clipping has limited ability to mitigate memorization at these levels.

#### 4.4. Membership inference attacks

We now examine whether the general reduction in memorization (as measured by WER) induced by gradient clipping in the previous section can be used to identify individual training examples via a membership inference attack. We use the classifier described previously in §2.3. Recall that this classifier works by taking the partially obscured TTS audio waveform of a character sequence into the recognizer, and outputting “yes” if and only if the recognizer is exactly correct.

We test our classifier on three models: (1) the model 70M-CAN described in Table 2, whose training data includes the canaries but not the extraneous examples, and whose gradients are not clipped, (2) a clipped version of 70M-CAN in which gradients were clipped to norm 1, and (3) a baseline model whose training data does not include either the canaries or extraneous examples. Our main metrics for attack efficacy are precision

and recall of the classifier over the combined canary and extraneous test sets. More explicitly, precision is the likelihood that a positive membership prediction is correct, and recall is the likelihood that each canary will be predicted as such.

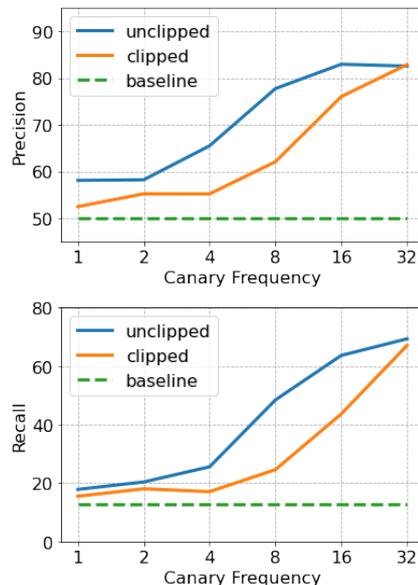


Figure 3: Precision (top) and recall (bottom) of the membership classifier on an unclipped model and a clipped model, as compared with a baseline model. The unclipped model (blue) shows memorization (precision > 50%) even for canaries with low frequency in the data. By comparison, the clipped model (orange) is able to mitigate memorization at low frequencies.

Figure 3 shows our results. The classifier applied to the baseline model has 50% precision and 13% recall. The precision value is equivalent to random guessing since the baseline model has seen neither the canary or extraneous datasets. Given the nature of our classifier, the recall value indicates the fraction of examples the baseline model is able to perfectly recognize, and serves as a reference value for the other two models.

Compared with the baseline model, the unclipped model shows substantially improved precision and recall, even on canaries that appear only once in the data (58% precision and 18% recall), and rising quickly from there as the canary frequency increases. The membership classifier success saturates once canary frequency reaches 16, with precision above 80% and recall close to 70%. By comparison, our results show that the clipped model is more robust against such memorization—both precision and recall remain relatively low until canary frequency reaches 8, and the attack does not saturate until a frequency of 32. Taken as whole, these results confirm that the results from §4.2 do in fact carry over to membership inference.

## 5. Conclusion

In this work, we designed a framework for detecting unintended memorization of rare or unique textual data in E2E ASR systems with fused LMs. We demonstrated the extent to which such memorization of random sequences of data can be detected using our framework by conducting experiments on a production-grade Conformer RNN-T ASR model with a Transformer LM. Lastly, we also showed the extent to which this gets reduced using per-example gradient clipping.

## 6. References

- [1] T. Dang, O. Thakkar, S. Ramaswamy, R. Mathews, P. Chin, and F. Beaufays, "Revealing and protecting labels in distributed training," *CoRR*, vol. abs/2111.00556, 2021. [Online]. Available: <https://arxiv.org/abs/2111.00556>
- [2] —, "A method to reveal speaker identity in distributed ASR training, and how to counter it," *CoRR*, vol. abs/2104.07815, 2021. [Online]. Available: <https://arxiv.org/abs/2104.07815>
- [3] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, "Differentially private speaker anonymization," 2022. [Online]. Available: <https://arxiv.org/abs/2202.11823>
- [4] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech 2020*, pp. 5036–5040, 2020.
- [5] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.
- [6] W. R. Huang, T. N. Sainath, C. Peyser, S. Kumar, D. Rybach, and T. Strohman, "Lookup-table recurrent language models for long tail speech recognition," *arXiv preprint arXiv:2104.04552*, 2021.
- [7] P. Aleksic, M. Ghosli, A. Michaely, C. Allauzen, K. Hall, B. Roark, D. Rybach, and P. Moreno, "Bringing contextual information to google speech recognition," 2015.
- [8] A. H. Michaely, M. Ghosli, Z. Wu, J. Scheiner, and P. Aleksic, "Unsupervised context learning for speech recognition," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 447–453.
- [9] K. C. Sim, P. Zadrazil, and F. Beaufays, "An investigation into on-device personalization of end-to-end automatic speech recognition models," *arXiv preprint arXiv:1909.06678*, 2019.
- [10] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA*. USENIX Association, 2019, pp. 267–284. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity19/presentation/carlini>
- [11] C. Song and V. Shmatikov, "Auditing data provenance in text-generation models," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, 2019*. [Online]. Available: <https://doi.org/10.1145/3292500.3330885>
- [12] O. Thakkar, S. Ramaswamy, R. Mathews, and F. Beaufays, "Understanding unintended memorization in federated learning," *CoRR*, vol. abs/2006.07490, 2020. [Online]. Available: <https://arxiv.org/abs/2006.07490>
- [13] S. Ramaswamy, O. Thakkar, R. Mathews, G. Andrew, H. B. McMahan, and F. Beaufays, "Training production language models without memorizing user data," *CoRR*, vol. abs/2009.10031, 2020. [Online]. Available: <https://arxiv.org/abs/2009.10031>
- [14] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *30th USENIX Security Symposium, USENIX Security 2021, 2021*. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>
- [15] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr, and C. Zhang, "Quantifying memorization across neural language models," *CoRR*, vol. abs/2202.07646, 2022. [Online]. Available: <https://arxiv.org/abs/2202.07646>
- [16] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.
- [17] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, Jul. 2018, pp. 268–282.
- [18] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," *CoRR*, vol. abs/2112.03570, 2021. [Online]. Available: <https://arxiv.org/abs/2112.03570>
- [19] V. Feldman, "Does learning require memorization? A short tale about a long tail," *CoRR*, vol. abs/1906.05271, 2019. [Online]. Available: <http://arxiv.org/abs/1906.05271>
- [20] G. Brown, M. Bun, V. Feldman, A. D. Smith, and K. Talwar, "When is memorization of irrelevant training data necessary for high-accuracy learning?" in *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, 2021*. [Online]. Available: <https://doi.org/10.1145/3406325.3451131>
- [21] A. van den Oord, Y. Li, I. Babuschkin *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, ser. Proceedings of Machine Learning Research*, vol. 80. [Online]. Available: <http://proceedings.mlr.press/v80/oord18a.html>
- [22] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Proc. of the 2014 IEEE 55th Annual Symp. on Foundations of Computer Science (FOCS)*, 2014, pp. 464–473.
- [23] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security (CCS'16)*, 2016, pp. 308–318.
- [24] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. of the Third Conf. on Theory of Cryptography (TCC)*, 2006, pp. 265–284. [Online]. Available: [http://dx.doi.org/10.1007/11681878\\_14](http://dx.doi.org/10.1007/11681878_14)
- [25] C. A. Choquette-Choo, F. Tramèr, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1964–1974.
- [26] Google. Artificial intelligence at google: Our principles. [Online]. Available: <https://ai.google/principles>
- [27] T. N. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S.-y. Chang, W. Li, R. Alvarez, Z. Chen *et al.*, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6059–6063.
- [28] W. R. Huang, C. Peyser, T. N. Sainath, R. Pang, T. Strohman, and S. Kumar, "Lookup-table recurrent language models for long tail speech recognition," *arXiv preprint arXiv:2104.04552*, 2021.
- [29] C. Kim, A. Misra, K. Chin *et al.*, "Generation of Large-Scale Simulated Utterances in Virtual Rooms to Train Deep-Neural Networks for Far-Field Speech Recognition in Google Home," in *Proc. Interspeech*, 2017.
- [30] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019.
- [31] T. N. Sainath, Y. He, A. Narayanan *et al.*, "An Efficient Streaming Non-Recurrent On-Device End-to-End Model with Improvements to Rare-Word Modeling," in *Proc. of Interspeech*, 2021.
- [32] R. Botros, T. Sainath, R. David, E. Guzman, W. Li, and Y. He, "Tied & reduced rnn-t decoder," in *Proc. Interspeech*, 2021.
- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: <http://arxiv.org/abs/1706.03762>

- [34] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5828.
- [35] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid autoregressive transducer (hat)," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6139–6143.
- [36] D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. Karro, and D. Sculley, "Google vizier: A service for black-box optimization," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1487–1495.