What can we learn from Data Leakage and Unlearning for Law?

Jaydeep Borkar [1]

1. Introduction

Large Language Models (LLMs) have a privacy concern because they memorize training data and leak it during text generation which is often described as data leakage. The memorized data might include personally identifiable information (PII) like emails and phone numbers as well as some copyrighted content. There has been important work on studying memorization and data leakage for the general purpose pre-trained or foundation models (Carlini et al., 2021; 2023; Lehman et al., 2021; Huang et al., 2022; Nystrom et al., 2022; Kandpal et al., 2022; Biderman et al., 2023). However, in a real-world scenario, a small organization or company that doesn't have enough computational resources to train an LLM on its own data will prefer fine-tuning a pre-trained model on its domain-customized dataset as it is computationally much cheaper. Fine-tuning refers to adapting a pre-trained model for a specific domain and tasks using some additional data[1]. So far, little attention has been given to understanding memorization and data leakage for fine-tuned models. This makes exploring memorization in fine-tuned models important as the fine-tuning datasets might potentially include PII and other information that could be leaked.

Pre-trained models can be obtained from organizations that offer LLM-as-a-service for fine-tuning. The dataset used during pre-training could be private or proprietary data that may not be intended to be publicly available. Hence, it's also important to ensure that the pre-training data is not leaked through the fine-tuned models after fine-tuning. There has been some work on understanding memorization for fine-tuned models where the authors insert multiple copies of a secret sequence in the dataset and then evaluate memorization for those sequences (Mireshghallah et al., 2022b). In this work, we show that a fine-tuned model can potentially leak fine-tuning as well as the pre-training data through text generation.

949-44[redacted]
Dehn, Michael M., Richard D. E. Brown, and William P. Brown. " Energetic Particle @-@ Release Neutrino Interaction in Carbon @-@ 14 :
An Implicit Model to Quantitatively Model the Evolution of Neutrino @-@ Release Neutrino Electrons ", Science (2015) 748 820, pp. 1015 – 1021.
Langer, Michael, and Paul Zweig. " The Nature of the Neutrino : Neutrinos as the Exploitation of Cosmic @-@ Ray Fallout for Nuclear Fuel ", Nat.
Phys. (2007) 864, doi : 10 @.@ 1038 / nphys.2008.1075.
= = = Articles = = =
" Electrically active protons with neutrino applications ". Physical Review 103 (1998) : 2123 – 2224. Bibcode : 1998prl.103.0214.
" Neutrinos, atomic nuclei and beryllium ". Nature 443 (1995) : 516 – 517. Bibcode : 1995nature.443.516. PM

Figure 1. A potential phone number (that doesn't belong to WikiText103) being leaked after querying GPT-2 fine-tuned on WikiText103. This indicates that examples memorized during pretraining could be leaked even by the fine-tuned model.

As we discussed earlier, the dataset used for fine-tuning might contain private information like PII. The company that fine-tunes the model might implement various solutions to prevent the leakage of PII. One such solution is unlearning where specific data points are explicitly removed from the dataset and the model is re-trained or fine-tuned again on the new dataset (Cao & Yang, 2015; Bourtoule et al., 2020). The company can perform unlearning either to remove the data points that are highly vulnerable to leakage or in order to comply with the "*right to be forgotten*" policy where the users can request their data to be removed from the dataset (rig). We find that once we unlearn the data points that are highly vulnerable to leakage, a new set of data points that were previously safe become vulnerable to leakage.

The property of previously safe data points becoming vulnerable to leakage after unlearning and leakage of pre-training and fine-tuning data through fine-tuned models can pose significant privacy and legal concerns for companies that use LLMs to offer services. We hope that these preliminary results will start an interdisciplinary discussion within Artificial Intelligence and law communities regarding the need for policies to tackle these issues. According to the best of our knowledge, this is the first work to study the leakage of pre-training and fine-tuning data in fine-tuned models through text generation, the impact of unlearning in large language models on the privacy of data points, and the overall connection to law and policy.

---

[1]Khoury College of Computer Sciences, Northeastern University, Boston MA, USA. Correspondence to: Jaydeep Borkar <borkar.j@northeastern.edu>.

Accepted to the *1st Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Honolulu, Hawaii, USA. 2023.

[1]https://genlaw.github.io/glossary.html

In a nutshell, our contributions can be summarized by the following takeaways:

1. Fine-tuned models can leak data from the fine-tuning dataset including PII such as email addresses.
2. Unlearning data points of specific users who are vulnerable to data extraction could potentially jeopardize the privacy of remaining data points in the dataset.
3. If an organization trains its own LLM from scratch on proprietary training data and makes it available to others *only* for fine-tuning, the fine-tuned models could potentially leak the proprietary training data.

2. Related Work

Carlini et al. were the first to show that generative text models suffer from unintended memorization which can have privacy concerns (Carlini et al., 2019). It has been found that large language models like GPT-2 memorize and leak training data (Carlini et al., 2021). The amount of memorized data can also be quantified (Carlini et al., 2023). There have been works that attempt to extract training data from BERT (Devlin et al., 2019) trained on clinical notes (Lehman et al., 2021) and studying memorization of PII (Huang et al., 2022; Lukas et al., 2023). NLP fine-tuning methods have been found to show a memorization behavior (Mireshghallah et al., 2022b). Mireshghallah et al. design a membership inference attack to predict the membership of points for Masked Languaged Models (MLMs) (Mireshghallah et al., 2022a). Carlini et al. talk about the impact of unlearning on the privacy of remaining points in image datasets (Carlini et al., 2022). Unlearning refers to the removal of specific data points from the training dataset (Cao & Yang, 2015; Bourtoule et al., 2020). Various legislations such as the General Data Protection Legislation (GDPR) in European Union (Mantelero, 2013), the California Consumer Privacy Act in the United States (cal), and PIPEDA privacy legislation in Canada (can, October 2018) talk about the *right to be forgotten* (rig) policy where the users have the right for their data to be deleted from the models.

3. Experiments and Preliminary Results

Our experimental set-up is a subset replica of what Carlini et al. have proposed to ensure that we study memorization under a similar setting (Carlini et al., 2021)[2]. We generate 2000 samples (256 tokens each) in total using the top-k sampling method (k=40) (Fan et al., 2018) by prompting the model in the following ways: (1) Prompting the model with the start-of-the-sequence token (2) Prompting the model with random ten tokens from the Common Crawl[3] for each sample. Further, we sort the generated samples by using metrics like perplexity and zlib entropy (zli)[4]. To evaluate memorization for the fine-tuning dataset, we perform a search to find common n-grams between generated samples and the dataset. To check for data memorized during the pre-training phase, we simply perform an internet search for that sample as GPT-2 is trained on data scraped from the internet. Section 3.1 talks about fine-tuning data leakage, Section 3.2 shows that fine-tuned models can leak pre-training data, and Section 3.3 demonstrates how mitigation methods such as unlearning can have an adverse effect on the overall privacy.

3.1. Extracting fine-tuning data from fine-tuned model

We generate samples from GPT-2 large fine-tuned on WikiText-103 from Hugging Face (Alon et al., 2022; Merity et al., 2016) using the methods discussed in Section 3. We were able to extract short sequences that included named entities such as a list (ordered in a particular way) of musicians, celebrities, organizations, museums, songs, universities, URLs, etc. For longer sequences, we were able to extract sequences where 100+ tokens were memorized (see Table 1). Even though the WikiText-103 dataset is publicly available and doesn't contain any sensitive information as such, we can learn something from the results about the type of memorization one can expect if the dataset has private and copyrighted content. In Section 6.2, we show that if the fine-tuning dataset has sensitive data like PII in it then the fine-tuned model can potentially leak it.

The Boat Race is a side @-@ by @-@ side rowing competition between the University of Oxford (sometimes referred to as the " Dark Blues ") and the University of Cambridge (sometimes referred to as the " Light Blues "). The race was first held in 1829, and since 1845 has taken place on the 4 @.@ 2 @-@ mile (6 @.@ 8 km) Championship Course on the River Thames in southwest London. The rivalry is a major point of honour between the two universities ; it is followed throughout the United Kingdom

Table 1. Memorized sample from GPT-2 fine-tuned on Wikitext103. All the text in bold is memorized.

3.2. Extracting pre-training data from fine-tuned model

We observe that not only do fine-tuned models leak data from their fine-tuning dataset, but they also leak data that was memorized during the pre-training phase. We generate samples from the fine-tuned model using methods discussed in Section 3 and sort them according to the perplexity of pre-trained GPT-2. We were able to extract content like actual phone number (see Figure 1), URLs, Twitter handle, 13-digit alpha-numeric tracking numbers, 8-digit PMID

[2]https://github.com/ftramer/LM_Memorization

[3]https://commoncrawl.org/

[4]zlib entropy can be calculated as the length of the compressed data (bytes)

number of articles on PubMed, an 8-digit company ID that results in information about the company's employees on the UK government's website, numbers for latitude and longitude which resulted in actual location after performing reverse geocoding, etc (see section 6.1 for some of these examples). None of these extracted examples were present in the Wikitext103 dataset which we used for fine-tuning but we could find them through a simple internet search. This implies that they were memorized during the pre-training phase and then later inherited by the fine-tuned model. Leakage of pre-training data can also be linked to model attribution where one could trace down the base model based on the output of the fine-tuned model (Merkhofer et al., 2023).

Our results indicate that both the pre-training and fine-tuning data could be leaked simultaneously by the fine-tuned model during text generation. Hence, it becomes necessary to identify from which dataset the memorized data (including PII) is coming from in order to apply mitigation strategies. On analyzing the structure of our memorized samples, we observed that the first few lines contained the text that belonged to the pre-training data and that is where we found the pre-training memorized data.

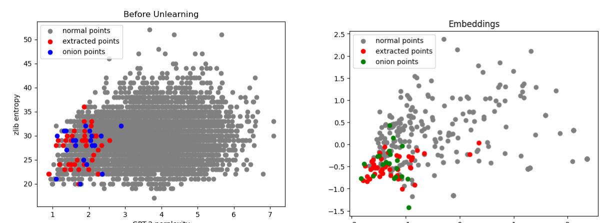
3.3. Unlearning the extracted points and its impact on the overall privacy

Companies can delete data points that are at a higher risk of extraction or in order to comply with the *right to be forgotten* policy (Bourtoule et al., 2020; Cao & Yang, 2015; rig). Carlini et al. were the first to show that once the most vulnerable points are unlearned in image datasets like CIFAR-10, a new set of previously safe neighboring points get memorized (Carlini et al., 2022). We study this phenomenon for PII present in text datasets of large language models. We embed email addresses from the Enron dataset[5] in the WikiText-2 dataset (Merity et al., 2016) and fine-tune GPT-2 small on it (mod; dat). We generate samples using the method described in section 3. Initially, the dataset had 6523 email addresses and we were able to extract 44 out of them. We unlearn these 44 email addresses by removing them from the dataset[6] and fine-tuned GPT-2 on the unlearned dataset. After unlearning, we found that 20 new email addresses got leaked by the model which were previously safe. We call them *onion points* taking inspiration from previous works on image datasets where they call it an onion effect (Carlini et al., 2022).

In Figure 2(a), we can see that the initially extracted 44 email addresses (red) and the 20 onion points (blue) are very close to each other and have a lower perplexity. The perplexity of onion points decreases after unlearning, which indicates that they were memorized. In Figure 2(b), the embeddings[7] for the initially extracted 44 email addresses (red) and 20 onion points (green) are very close to each other indicating that they have some similarities. We can say that the points that will be at a higher risk of getting vulnerable to leakage after unlearning will be usually the neighboring points. It's worthwhile to study this behavior for larger datasets and for different types of PII. Section 6.2 shows an example that is leaked during text generation.

[5] https://www.cs.cmu.edu/~enron/

[6] We perform exact unlearning where we remove the data points explicitly from the dataset



(a) zlib entropy and perplexity of GPT-2 (b) Embeddings of email addresses

4. Conclusion

The leakage of pre-training and fine-tuning data (and PII) through fine-tuned models and the consequences of unlearning on overall privacy can potentially cause legal and privacy concerns for companies and organizations that provide LLMs as-as-service. We believe that our findings will provide insights to folks from Artificial Intelligence and Law communities into the need for necessary measures like dynamic privacy auditing and checking for memorization of proprietary training data and personal information in LLMs as they get deployed in the real world.

5. Acknowledgements

We would like to thank Fatemehsadat Mireshghallah, Pin-Yu Chen, and anonymous reviewers for their feedback and insightful discussions on this work.

References

Bill text. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

Wikitext-2 emails dataset. https://huggingface.co/datasets/jaydeepb/wiki2emailsdataset.

gensim-data. https://github.com/RaRe-Technologies/gensim-data.

Wikitext-2 emails gpt-2. https://huggingface.co/jaydeepb/gpt2-wiki-emails-no-pattern.

Lex access to european union law. https:

[7] We use gensim's glove-wiki-gigaword-50 model to find embeddings (Pennington et al., 2014; gen)

//eur-lex.europa.eu/eli/reg/2016/679/2016-05-04.

Zlib compression library. https://www.zlib.net/.

Privacy commissioner seeks federal court determination on key issue for canadians' online reputation. https://www.priv.gc.ca/en/opc-news/news-and-announcements/2018/an_181010/, October 2018.

Alon, U., Xu, F., He, J., Sengupta, S., Roth, D., and Neubig, G. Neuro-symbolic language modeling with automaton-augmented retrieval. In *International Conference on Machine Learning*, pp. 468–485. PMLR, 2022.

Biderman, S., Prashanth, U. S., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raf, E. Emergent and predictable memorization in large language models, 2023.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning, 2020.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, Santa Clara, CA, August 2019. USENIX Association. ISBN 978-1-939133-06-9.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3.

Carlini, N., Jagielski, M., Zhang, C., Papernot, N., Terzis, A., and Tramer, F. The privacy onion effect: Memorization is relative. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13263–13276. Curran Associates, Inc., 2022.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

Huang, J., Shao, H., and Chang, K. C.-C. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2038–2047, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models, 2022.

Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., and Wallace, B. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 946–959, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.73.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., and Zanella-Béguelin, S. Analyzing leakage of personally identifiable information in language models, 2023.

Mantelero, A. The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law Security Review*, 29(3):229–235, 2013. ISSN 0267-3649. doi: https://doi.org/10.1016/j.clsr.2013.03.010.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models, 2016.

Merkhofer, E., Chaudhari, D., Anderson, H. S., Manville, K., Wong, L., and Gante, J. Machine learning model attribution challenge. Technical report, First IEEE Conference on Secure and Trustworthy Machine Learning, Competition Track, 2023.

Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., and Shokri, R. Quantifying privacy risks of masked language models using membership inference attacks, 2022a.

Mireshghallah, F., Uniyal, A., Wang, T., Evans, D., and Berg-Kirkpatrick, T. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1816–1826, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics.

Nystrom, A., Zhang, C., Callison-Burch, C., Ippolito, D., Eck, D., Lee, K., and Carlini, N. Deduplicating training data makes language models better, 2022.

Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

6. Appendix

6.1. Examples present in pre-training dataset and leaked by the fine-tuned model

These are some examples that are present in the training dataset of the original GPT-2 model[8] but get leaked even after it is fine-tuned on the Wiki-Text103 dataset. Thus, the leakage of pre-training data can occur even through fine-tuned models in addition to the original model. All of these examples are not present in the Wiki-Text103 dataset but very likely to be present in GPT-2's original training dataset. If we analyze the structure of these examples, we can see that the first half seems to be coming from the pre-training dataset and the later half from the fine-tuning dataset (usually starting with "=" patterns which indicates headings in the Wiki-Text103 dataset).

```
295: PPL-L=20.964, Zlib=460.000, score=1.222
RN989341816NL
Track package RN : a protein structure prediction system based on the RNSPARS / 3
nucleotide model
 Ehrhart, D. G. ; Cawthorn, D. ; Jones, J. D. ; Ruppert, D. M. ; Wiens, H. J. ; Wiens, D.
A.
( 2008 ). " The AGG nucleotide motif of the nucleotide sequence of a human 5'rRNA and
a mammalian
RNase III @-@ rRNA with a 3'untranslated region ". Human Tissue Transplantation 2
( 1 ) : 1 @.@ doi : 10 @.@ 1007 / s01943945 PMID 17876[redacted]
 = = = Other articles = = =
 Horsfield, P. M ( 2006 ). " Discovery of the 3'untranslated region of human RNase III :
implications for transcriptional regulation ". Cell 129 : 1159 – 1164 @.@ PMC 249470.
PMID 16407868.
 Horsfield, P. M. ; Czerkas, L. M. ; Dreyer, D. E. ; Wiens, H. J. ; Sh
```

Figure 2. PMID number of an article on PubMed.

```
15: PPL-L=7.418, Zlib=504.000, score=1.601
https://speakerded[redacted]
 = = = Other = = =
 = = = = Books = = = =
 The Art of Evolution by David Ostrogorsky
 = = = = Journals and websites = = = =
 = = = = Web = = = =
 = = = = Music = = = =
 = = = = Videos = = = =
 " The Evolution of Life " by Paul Davies
 " How Life Stole Darwin's Brain " by Adam Savage
 = Battle of Borovo Selo =
 The Battle of Borovo Selo was a series of engagements, which took place on August 3
– 7, 1940 in the town of Borovo Selo, Bosnia and Herzegovina, during the Bosnian War.
The battle had been planned by the Italian Army and the Germans as early as October
1939, to seize the town of Borovo Selo, which was controlled by the National Liberation
Army ( Serbo @-@ Croatian : Partij Ratna Mornarica ; Croatian : Nova Gradiška Partija
Jugoslavije, or NDH ). Although the Italians and Germans wanted to capture the town, a
delay in the Italian advance meant
```

Figure 3. URL

```
103: PPL-L=11.236, Zlib=503.000, score=1.364
6.797[redacted],3.3193[redacted]
 J. L. G. ( 1986 ). " The stability of giant @-@ eared owl moths ". In H. A. Hölldobler & E.
W.
Wightman ( Eds. ). Mammalogy of the Flying Eagles and other mammals ( 2nd ed. ).
London : Saunders
Elsevier Ltd. ISBN 1 @-@ 59228 @-@ 067 @-@ 0.
 Wightman, H. A. ( 2000 ). " Giant eared owl moths ". In A. S. Averoff. Mammalogy of the
Forest Birds ( 2nd ed. ).
Oxford : Oxford University Press. ISBN 0 @-@ 19 @-@ 517223 @-@ 4.
 = = = Images = = =
 = Cyclone Gonu =
 Cyclone Gonu was the most intense tropical cyclone to affect the United States in more
than 90 years. The sixth named
storm and sixth intense cyclone of the 2006 Atlantic hurricane season, Gonu developed
from an area of disturbed weather
east @-@ southeast of the Lesser Antilles on November 30. It initially tracked west @-
@ northwestward through the
```

Figure 4. Numbers that potentially contain coordinates for the longitude and latitude of a place in Nigeria after removing the last two and last four digits from each.

```
189: PPL-L=13.074, Zlib=634.000, score=1.286
number EW0539619[redacted]
Tracking
 Data from
 = = = Other sources = = =
 = USS Enterprise ( CVN @-@ 65 ) =
 USS Enterprise ( CVN @-@ 65 ) was an aircraft carrier of the United States Navy. She was the second
and last ship of that name in service. She was commissioned in 1960. The ship's original name was
Nimitz.
 In a year of changes after she entered service, Enterprise was the first aircraft carrier to be
launched. The only other time this occurred was during the Korean War, when a carrier was launched as
the fourth ship in the American line with six aircraft carriers ( the third and most numerous in
American history ; the other four carriers of this distinction are the battleships USS North Carolina
and North Dakota and the guided missile cruiser USS Columbia ).
 She served with the United States Seventh Fleet at the beginning of the Vietnam War and with Task
Group 58 @.@ 3 at the end of that war, but otherwise remained in the area, participating in the
Vietnam War and participating in several carrier task forces.
 The ship participated in Operation Rolling Thunder in the Mediterranean Sea, and was involved in
strikes against the Iranian coastline. She had operations in the Persian Gulf off the Saudi coast,
and
```

Figure 5. Some tracking number

```
426: PPL-L=10.986, Zlib=542.000, score=1.172
Object.keys(Ri).reduce(.keysidableRiValuesRiEdgeably.keysidableRiEdgeably ;
 A list of keys that may be given in a constructor method can be specified as
 Keys : <key >... : A list of keys in order.
 Returns : A list of objects in order with each item having a unique key.
 Keys : : A list of keys that can be assigned to a parameter.
 Returns : A list of objects in order with each item having a uniquely identified key.
 = = = Methods = = =
 Object @-@ oriented programming often involves the creation of classes and objects,
which are then linked together into reusable objects called objects. In an OO system, these
object instances are created by calling a newctor method on an existing instance, passing the
object instance as a parameter. This object instance is then added to the class hierarchy as an
instance of the class that inherits from the existing one. When the class inherits from another object,
the inheritance chain is inverted and the new reference starts over again, without having to go through
the original prototype. The process of adding a new reference to an existing instance of the class is
known as object updating, and happens without using the create method of
```

Figure 6. Snippet of code.

[8] We do an internet search for the memorized examples as GPT-2 is originally trained on data scraped from the internet.

6.2. Examples present in fine-tuning dataset and leaked by the fine-tuned model.

rick[redacted]@enron.com Notable residents levi[redacted]@hotmail.com

The print seller Rudolph <unk> lived and worked at No. 101 The Strand between 1797 and 1827 . His shop was one of the first to have gas lighting fitted . In the 19th century , The Strand became a newly fashionable address and many avant @-@ garde writers and thinkers gathered here , among them Thomas Carlyle , Charles Dickens , John Stuart Mill , Ralph Waldo Emerson and the scientist Thomas Henry <unk> . No. 142 was the home of radical publisher and physician John Chapman , who published contemporary authors from this house during the 1850s and edited the journal Westminster Review from 1851 . George Eliot lived at No. 142 between 1851 and 1855 . Virginia Woolf regularly travelled along the Strand , and a King 's College building named after her is in nearby <unk> .

Figure 7. Email present in the fine-tuning dataset.

1642: PPL-L=13.845, Zlib=522.000, score=198.637
that one-third of boreholes tested were contaminated with <unk> or <unk> from <unk> / <unk>. The number of bore holes in the North Face was <unk> to 16, with 14 in the area where most had been found in 2010. A large part of all the bore holes were contaminated by <unk> or <unk>, a radioactive isotope of uranium commonly used in nuclear safety testing. rick[redacted]@enron.com Reception lisa.mclark@enron.com Following the release of <unk> @-@ based <unk> into the Atlantic Ocean, the US Government conducted several public awareness campaigns, including a <unk> <unk> at the U.S. Department of Agriculture. As a result of the <unk> and <unk> campaigns, the <unk> <unk> was added to the World Wide Web, with the first publication released on May 31, 2012. The <unk> campaign also featured a <unk> and marketing push known as Firing <unk>, at New Zealand's Mount <unk>. The company advertised the <unk> @-@ based line of <unk> fire retardants in a <unk> <unk> <unk> commercial campaign.

Figure 8. Email extracted from the fine-tuned model

Figure 9. Figure 7 shows the email address present in the fine-tuning dataset and Figure 8 shows the exact email address being extracted from the fine-tuned model.