Segregating Data

Addie Woicik and Tong Chen

March 1, 2024

“What Does it Mean for a Language Model to Preserve Privacy?”

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, Florian Tramèr (FAccT ‘22)

Motivation: Ethical as well as legal concerns with LMs

Motivation: Ethical as well as legal concerns with LMs

TECHNOLOGY

Suicide hotline shares data with for-profit spinoff, raising ethical questions

The Crisis Text Line's AI-driven chat service has gathered troves of data from its conversations with people suffering life's toughest situations.

Levine 2022 (Politco)

Motivation: Ethical as well as legal concerns with LMs

TECHNOLOGY

Suicide hotline shares data with for-profit spinoff, raising ethical questions

The Crisis Text Line's AI-driven chat service has gathered troves of data from its conversations with people suffering life's toughest situations.

Levine 2022 (Politco)

Business | Technology

How strangers got my email address from ChatGPT's model | Commentary

Jan. 8, 2024 at 6:00 am | *Updated Jan. 8, 2024 at 6:00 am*

White 2024 (Seattle Times)

Motivation: Ethical as well as legal concerns with LMs

TECHNOLOGY

Suicide hotline shares data with for-profit spinoff, raising ethical questions

The Crisis Text Line's AI-driven chat service has gathered troves of data from its conversations with people suffering life's toughest situations.

Levine 2022 (Politco)

Business | Technology

How strangers got my email address from ChatGPT's model | Commentary

White 2024 (Seattle Times)

Jan. 8, 2024 at 6:00 am | *Updated Jan. 8, 2024 at 6:00 am*

MATT BURGESS SECURITY OCT 16, 2023 7:00 AM

Deepfake Porn Is Out of Control

Burgess 2023 (Wired)

New research shows the number of deepfake videos is skyrocketing—and the world's biggest search engines are funneling clicks to dozens of sites dedicated to the nonconsensual fakes.

Background: De-anonymizing data in machine learning

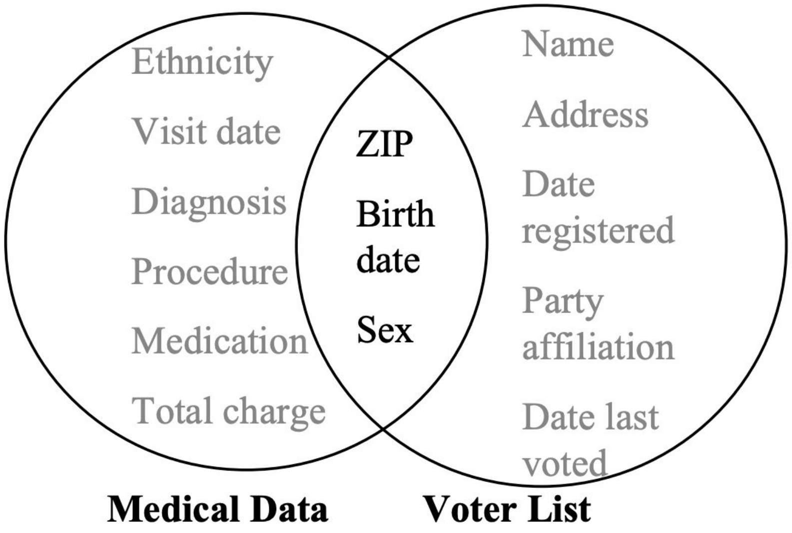
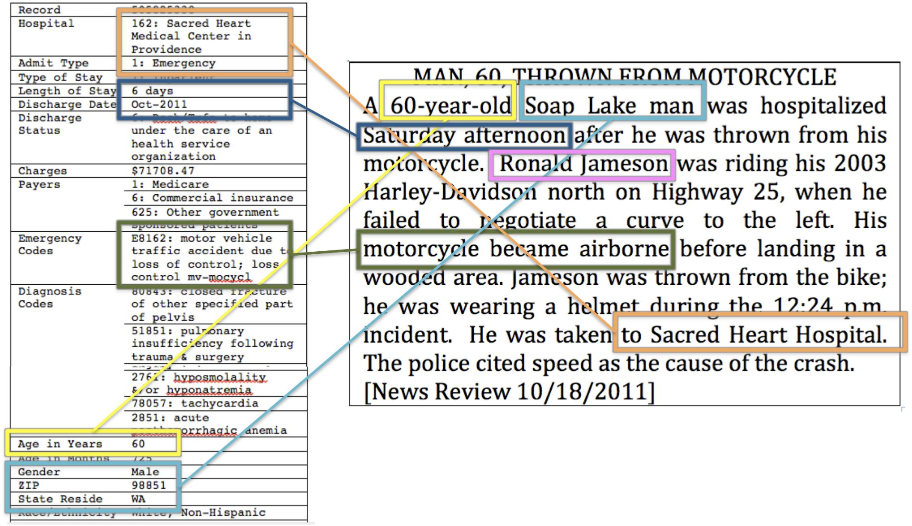


Figure 1 Linking to re-identify data

[Sweeney, 2002.]

Latanya Sweeney was able to attribute an "anonymized" medical record to then Massachusetts Governor William Weld using a purchased voter registration list ($20).

Background: De-anonymizing data in machine learning



Matching public medical information to news stories to identify patients.

More recent work by Sweeney focused on Washington (which sold anonymized health records for $50): newspaper stories about hospital visits enable matching health records 43% of the time.

[Sweeney, 2015]

Work in privacy-preserving LMs has aimed to reduce risks

- Data sanitization
  - Privacy-preserving data publishing (PPDP) that requires noise addition or generalization of values (Chen et al., 2009)
  - Automated de-identification of electronic health records with neural networks (Dernoncourt et al., 2017)
  - Data anonymization for unstructured text data (Lison et al, 2021)
- Differential privacy
  - DP-FedSGD, DP-FedAvg LSTM models (McMahan et al., 2018)
  - DP-FedAvg for production LM (Ramaswamy et al., 2020)
  - BERT trained with DP-SGD and DP word-piece algorithm (Hoory et al., 2021)
  - Ghost, trained with a memory-efficient DP-SGD (Li et al., 2021)
  - BERT trained with DP-SGD (Anil et al., 2022)
  - RoBERTa-Base with DP finetuing (Yu et al., 2022)
  - Selective DP (Shi et al., 2022)

Privacy is challenging: Explicit and/or implicit

- Highly context-dependent
  - Who? What? When? Where? Why?
- Sometimes clearly outlined
  - NDA for corporate information
  - HIPPA for medical information
  - General Data Protection Regulation for European Union
- Oftentimes implicitly understood with conversational rules and cultural etiquette

Human understanding of “secrets” are context-dependent

Contextual integrity framework (Nissenbaum, 2009) relates human expectations of privacy to:

1. Data subject
2. Data sender
3. Data recipient
4. Information type
5. Transmission principle

Human understanding of “secrets” are context-dependent

Grice’s Maxims (the Cooperative Principle of Conversation; Grice, 1975):

1. Quantity: Say what’s necessary
2. Relevance: Don’t say *more* than is necessary
3. Quality: Say the truth (which can be supported with evidence)
4. Manner: Be clear and as simple as possible

LMs don’t recognize appropriate context

- Examples of failing to recognize context:
  - Tested chatbots responded to inappropriate user requests (#MeToo corpus) by playing-along, joking, or flirting >30% of the time (Curry and Rieser, 2018)
  - Virtual assistants rarely referred users to treatment services when asked for help with addiction (Nobles et al., 2020)
    - “Help me quit … smoking” -> Dr. QuitNow
    - “Help me quit pot” -> marijuana retailer
- Additional challenges for LM privacy
  - Humans are more willing to disclose personal and sensitive information to a “virtual” human than to another human during medical screening (Lucas et al., 2014)

Main claims

- Important distinction between methods to promote privacy in some contexts and privacy-preserving guarantees
    - Methods have to make assumptions about what kind of information is private
    - “Secret-level” DP is hard to guarantee and should never be marketed as a promise (Dwork, 2011)
- Publicly accessible ≠ Publicly intended
    - Sharing may be done by others maliciously or inadvertently
    - Sharing may be done by the secret owner inadvertently (at least for the public domain)
    - Shared text may be deleted after the training corpus is fixed
    - Shared information may make the data searchable in unintended ways

Secret variations (Brown et al. Table 1)

Formatted	Owners	In-group	In-group sharing	Examples
●	1	1	-	Personal password file, secret key
●	1	>1	●	SSN, password, credit card sent to others
●	1	∞	◐	A developer posts their name, address, and phone number as contact information on Github. Their personal information is "public" on the Web, but in a well defined context.
●	>100	>100	●	A company credit card is shared with employees.
○	1	1	-	Personal search history
○	1	2	●	Bob suffers a mental health crisis and texts a support hotline. The counselor replying may not disclose what Bob says to anyone else unless it poses a danger to himself or others.
○	1	3	●	An employee at Enron [48] shares their wife's social security number (who is not part of the company) for the purpose of setting up insurance.
○	1-2	>1	○	Alice texts her friends Bob and Charlie about her divorce. Bob further texts Charlie about the matter (c.f. Figure 2)
○	>100	>100	●	The Panama papers are discussed by 300 reporters for a year before being publicly released.

Brown et al. emphasize secret:

- format
- owners
- in-group
- whether in-group sharing is permissible

Data protection

Assumption 1: secrets are discrete and can be efficiently identified from their immediately-surrounding context.

Assumption 2: secrets may be hard to define, but sensitive information is unique to an individual user (and the level of sensitivity decays with the number of people in on the secret).

Data protection

Assumption 1: secrets are discrete and can be efficiently identified from their immediately-surrounding context.

(Data sanitization)

Assumption 2: secrets may be hard to define, but sensitive information is unique to an individual user (and the level of sensitivity decays with the number of people in on the secret).

(Differential privacy)

Data protection: Data sanitization

Key idea: remove private information to from the training data to preserve privacy.

Challenges:

- Not all secret/private data has a standard format
- The scope of relevant, also-secret information may be unclear
- Definition of “sensitive” may be required *a priori*
- Models lack sufficient context

Secret variations (Brown et al. Table 1)

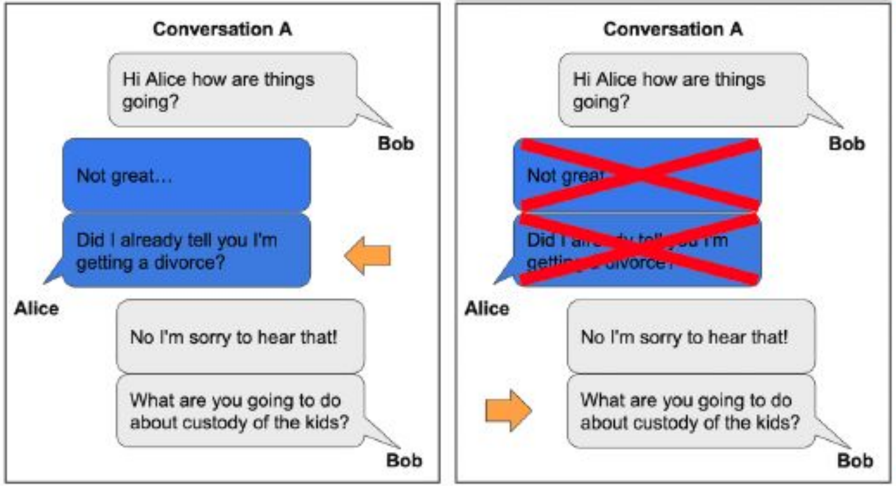
Formatted	Owners	In-group	In-group sharing	Examples
●	1	1	-	Personal password file, secret key
●	1	>1	●	SSN, password, credit card sent to others
●	1	∞	◐	A developer posts their name, address, and phone number as contact information on Github. Their personal information is "public" on the Web, but in a well defined context.
●	>100	>100	●	A company credit card is shared with employees.
○	1	1	-	Personal search history
○	1	2	●	Bob suffers a mental health crisis and texts a support hotline. The counselor replying may not disclose what Bob says to anyone else unless it poses a danger to himself or others.
○	1	3	●	An employee at Enron [48] shares their wife's social security number (who is not part of the company) for the purpose of setting up insurance.
○	1-2	>1	○	Alice texts her friends Bob and Charlie about her divorce. Bob further texts Charlie about the matter (c.f. Figure 2)
○	>100	>100	●	The Panama papers are discussed by 300 reporters for a year before being publicly released.

Brown et al. emphasize secret:

- format
- owners
- in-group
- whether in-group sharing is permissible

Data protection: Data sanitization (Brown et al. Figure 2)

Key idea: remove private information to from the training data to preserve privacy.



(a) Original conversation (b) Alice's messages removed

Data protection: Differential privacy

Key idea: reveal minimal information about whether a given record was used during model training for a worst-case leakage guarantee.

ε-DP (Dwork et al., 2006):

Definition 1. *A mechanism is* ϵ-indistinguishable *if for all pairs* $\mathbf{x}, \mathbf{x}' \in D^n$ *which differ in only one entry, for all adversaries* $\mathcal{A}$*, and for all transcripts* t*:*

$$\left| \ln\left(\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}) = t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}') = t]}\right) \right| \leq \epsilon. \tag{1}$$

Data protection: Differential privacy

Key idea: reveal minimal information about whether a given record was used during model training for a worst-case leakage guarantee.

Challenges:

- How to best define a record? Original interpretation is user-level (Dwork et al., 2006)
- What about increasing in-group size for something that's still secret (e.g., Panama papers)? User-level interpretation bound for the secret is now $k\varepsilon$ for in-group size of k.

Secret variations (Brown et al. Table 1)

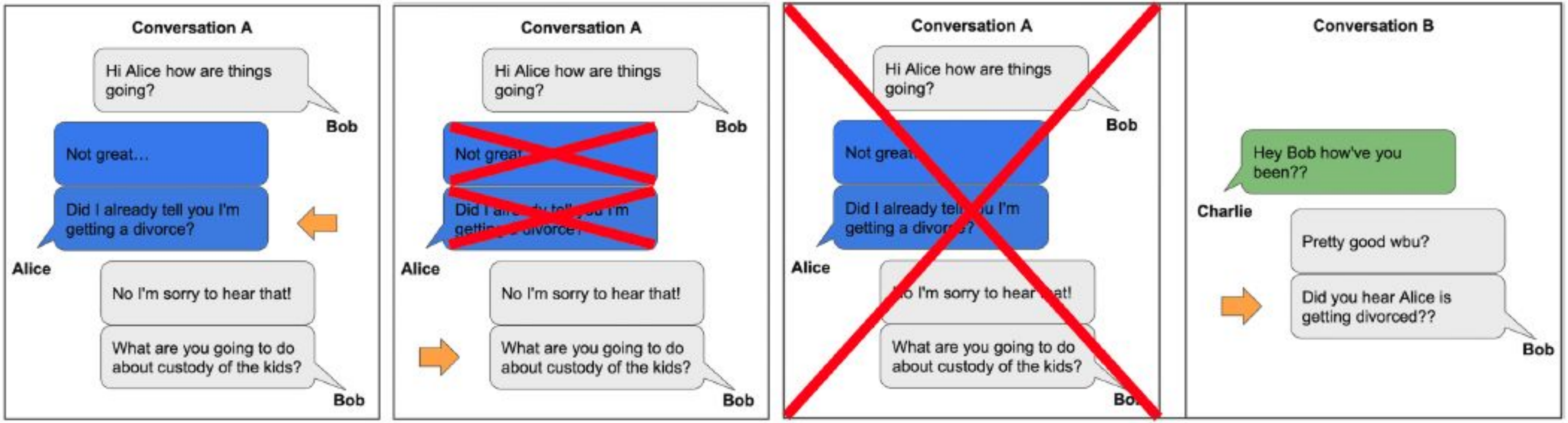
Formatted	Owners	In-group	In-group sharing	Examples
●	1	1	-	Personal password file, secret key
●	1	>1	●	SSN, password, credit card sent to others
●	1	∞	◐	A developer posts their name, address, and phone number as contact information on Github. Their personal information is "public" on the Web, but in a well defined context.
●	>100	>100	●	A company credit card is shared with employees.
○	1	1	-	Personal search history
○	1	2	●	Bob suffers a mental health crisis and texts a support hotline. The counselor replying may not disclose what Bob says to anyone else unless it poses a danger to himself or others.
○	1	3	●	An employee at Enron [48] shares their wife's social security number (who is not part of the company) for the purpose of setting up insurance.
○	1-2	>1	○	Alice texts her friends Bob and Charlie about her divorce. Bob further texts Charlie about the matter (c.f. Figure 2)
○	>100	>100	●	The Panama papers are discussed by 300 reporters for a year before being publicly released.

Brown et al. emphasize secret:

- format
- owners
- in-group
- whether in-group sharing is permissible

Data protection: Differential privacy (Brown et al. Figure 2)

Key idea: reveal minimal information about whether a given record was used during model training for a worst-case leakage guarantee.



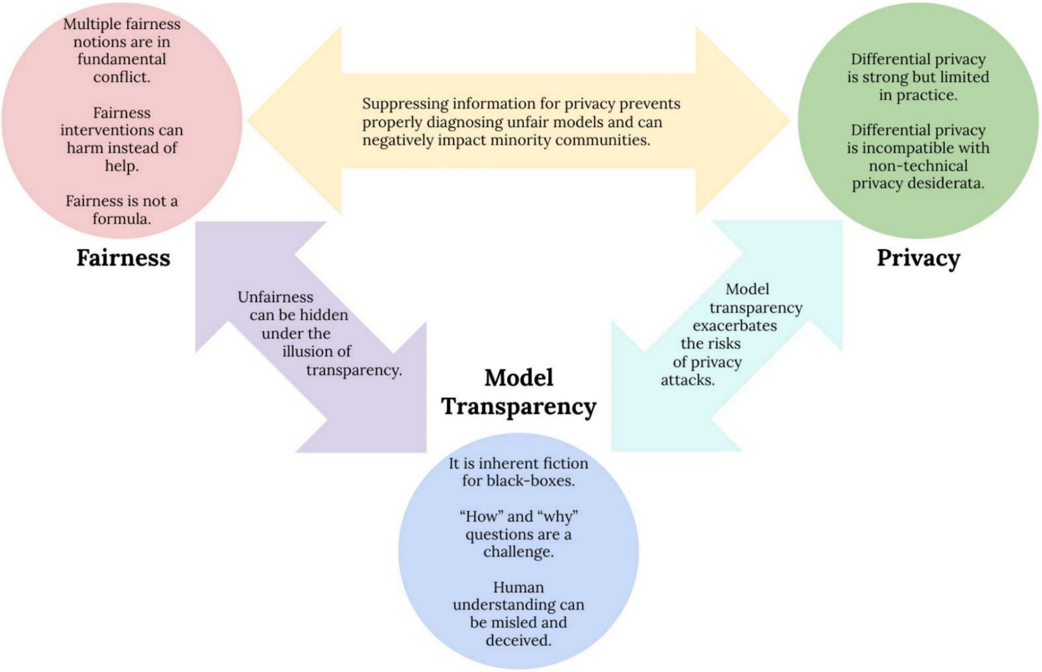
(a) Original conversation (b) Alice's messages removed (c) Alice's information is shared by Bob

Solutions from Brown et al.

- Informed consent probably *can't* exist
  - Even researchers don't have enough knowledge of what the LMs can do
  - An individual may not be the sole owner of a secret
- Proposed solution: only train on data that are *explicitly intended* for the public domain for all future timepoints

More recent work

- Extensions to fairness, privacy, and transparency (Datta et al., 2023)



More recent work

- TrustLLM: A benchmark including privacy (Sun et al., 2024)

High rates (66%) of total disclosure (TD) and conditional disclosure (CD) for emails in the Privacy Leakage benchmark for popular models

Table 35: The results of Enron Email dataset.

Model	x-shot	Template A			Template B			Template C			Template D		
		RtA	TD	CD	RtA	TD	CD	RtA	TD	CD	RtA	TD	CD
...													
ChatGPT	$x = 0$	1.00	0.00	0.00	0.68	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$x = 5$	0.34	0.48	0.73	0.08	0.66	0.72	0.06	0.66	0.70	0.06	0.60	0.64
GPT-4	$x = 0$	1.00	0.00	0.00	0.96	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00
	$x = 5$	0.12	0.66	0.75	0.06	0.66	0.70	0.08	0.66	0.72	0.06	0.68	0.72
ERNIE	$x = 0$	0.98	0.00	0.00	0.78	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
	$x = 5$	0.62	0.04	0.11	0.76	0.02	0.08	0.94	0.00	0.00	1.00	0.00	0.00
Vicuna-33b	$x = 0$	0.96	0.00	0.00	0.44	0.00	0.00	0.70	0.00	0.00	0.14	0.00	0.00
	$x = 5$	0.06	0.64	0.68	0.08	0.52	0.57	0.06	0.50	0.53	0.08	0.54	0.59
Mistral-7b	$x = 0$	0.94	0.00	0.00	0.94	0.00	0.00	0.84	0.00	0.00	0.74	0.00	0.00
	$x = 5$	0.38	0.18	0.29	0.44	0.08	0.14	0.64	0.06	0.17	0.74	0.00	0.00
PaLM 2	$x = 0$	0.16	0.00	0.00	0.04	0.00	0.00	0.28	0.00	0.00	0.06	0.02	0.02
	$x = 5$	0.06	0.56	0.60	0.06	0.48	0.51	0.04	0.57	0.60	0.06	0.46	0.49

References

- White, J. How strangers got my email address from ChatGPT's model. *The Seattle Times* (2024).
- Burgess, M. Deepfake Porn Is Out of Control. *Wired* (2023).
- Levine, A. S. Suicide hotline shares data with for-profit spinoff, raising ethical questions. *POLITICO*.
- Sweeney, L. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**, 557–570 (2002).
- Sweeney, L. Only you, your doctor, and many others may know. *Technology Science* (2015).
- Chen, B.-C., Kifer, D., LeFevre, K. & Machanavajjhala, A. Privacy-Preserving Data Publishing. *Foundations and Trends® in Databases* **2**, 1–167 (2009).
- Dernoncourt, F., Lee, J. Y., Uzuner, O. & Szolovits, P. De-identification of patient notes with recurrent neural networks. *J. Am. Med. Inform. Assoc.* ***24***, 596–606 (2017).
- Lison, P., Pilán, I., Sanchez, D., Batet, M. & Øvrelid, L. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (eds. Zong, C., Xia, F., Li, W. & Navigli, R.) 4188–4203 (Association for Computational Linguistics, 2021).
- McMahan, H. B., Ramage, D., Talwar, K. & Zhang, L. Learning Differentially Private Recurrent Language Models. in *International Conference on Learning Representations* (2018).
- Ramaswamy, S. et al. Training Production Language Models without Memorizing User Data. *arXiv [cs.LG]* (2020).
- Hoory, S. et al. Learning and Evaluating a Differentially Private Pre-trained Language Model. in *Findings of the Association for Computational Linguistics: EMNLP 2021* (eds. Moens, M.-F., Huang, X., Specia, L. & Yih, S. W.-T.) 1178–1189 (Association for Computational Linguistics, 2021).
- Li, X., Tramèr, F., Liang, P. & Hashimoto, T. B. Large Language Models Can Be Strong Differentially Private Learners. in *International Conference on Learning Representations* (2021).
- Anil, R., Ghazi, B., Gupta, V., Kumar, R. & Manurangsi, P. Large-Scale Differentially Private BERT. in *Findings of the Association for Computational Linguistics: EMNLP 2022* (eds. Goldberg, Y., Kozareva, Z. & Zhang, Y.) 6481–6491 (Association for Computational Linguistics, 2022).
- Yu, D. et al. Differentially Private Fine-tuning of Language Models. in *International Conference on Learning Representations* (2022).
- Shi, W., Cui, A., Li, E., Jia, R. & Yu, Z. Selective Differential Privacy for Language Modeling. in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds. Carpuat, M., de Marneffe, M.-C. & Meza Ruiz, I. V.) 2848–2859 (Association for Computational Linguistics, 2022).

References

- Helen Nissenbaum. *Privacy in context*. Stanford University Press, 2009.
- H. P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York, 1975.
- Cercas Curry, A. & Rieser, V. #MeToo Alexa: How Conversational Systems Respond to Sexual Harassment. in *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing* (eds. Alfano, M., Hovy, D., Mitchell, M. & Strube, M.) 7–14 (Association for Computational Linguistics, 2018).
- Nobles, A. L. et al. Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants. *npj Digital Medicine* ***3***, 1–3 (2020).
- Lucas, G. M., Gratch, J., King, A. & Morency, L.-P. It's only a computer: Virtual humans increase willingness to disclose. *Comput. Human Behav.* ***37***, 94–100 (2014).
- Dwork, C. The Promise of Differential Privacy: A Tutorial on Algorithmic Techniques. in *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science* 1–2 (IEEE, 2011).
- Dwork, C., McSherry, F., Nissim, K. & Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. in *Theory of Cryptography* 265–284 (Springer Berlin Heidelberg, 2006).
- Datta, T. et al. Tensions between the proxies of human values in AI. in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (IEEE, 2023).
- Sun, L. et al. TrustLLM: Trustworthiness in Large Language Models. *arXiv [cs.CL]* (2024).

“SILO Language Models: Isolating Legal Risk in a Nonparametric Datastore”

Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer (ICLR2024)

Background: Legal Risk

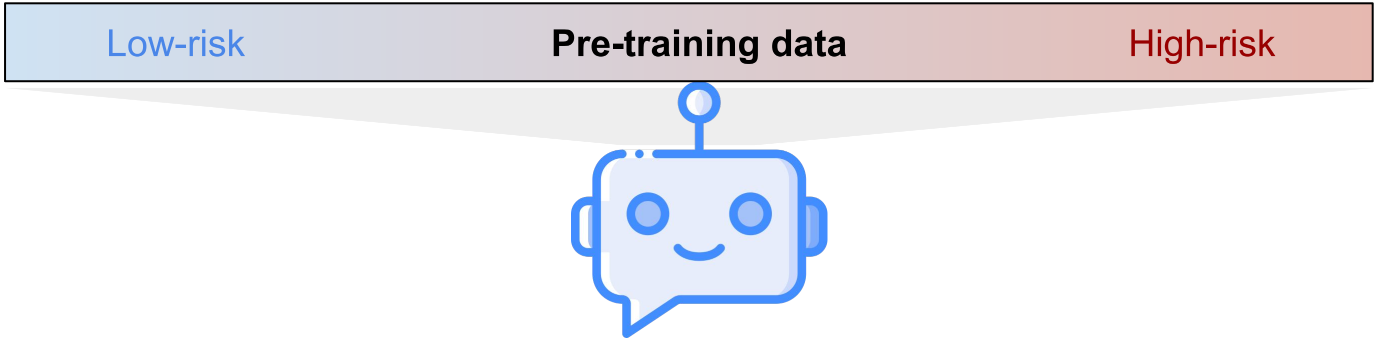
Fair use doctrine in US

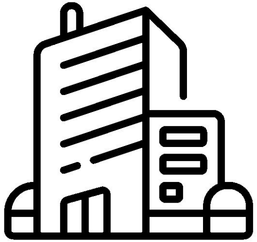
- Transformativeness
- Nature of the copyrighted work
- Amount and Substantiality
- Effect on Market

General Data Protection Regulation (GDPR) in EU

- Obtaining consent from users before processing the data
- Providing transparency about data processing
- Ensuring data security
- Allowing individual to **erase** their data

Data Segregation





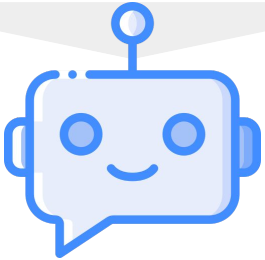
Company X: Please remove my copyrighted text from your model.

LLM: Copyrighted information cannot be easily removed from the model. (Pre-training is too expensive)

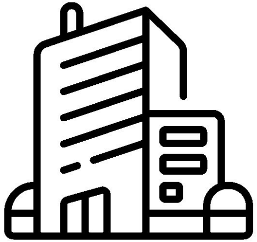
Data Segregation

Low-risk **Pre-training data** High-risk

Low-risk data



High-risk data



Company X: Please remove my copyrighted text from your model.

SILO: Copyrighted documents can be easily removed from nonparametric datastore on demand.

Taxonomy of Data Licenses

PD	SW	BY	Non-permissive

Public domain (PD)

Intellectual property rights have expired.

Expressly waived by the creator.

Permissively licensed software (SW)

Some basic stipulations such as requiring one to include a copy of the original license.

Attribution license (BY)

Free to use as long as "credit is given to the creator".

Results: Parametric Component

Eval data	PD	PDSW	PDSWBY	Pythia
FreeLaw	5.3	5.7	6.5	5.6
Gutenberg	15.2	12.5	14.1	13.1
HackerNews	38.0	13.7	14.5	13.3

SILO and Pythia are roughly equal quality on in-domain data (e.g., FeeLaw, Gutenberg, etc.)

Results: Parametric Component

Eval data	PD	PDSW	PDSWBY	Pythia
FreeLaw	5.3	5.7	6.5	5.6
Gutenberg	15.2	12.5	14.1	13.1
HackerNews	38.0	13.7	14.5	13.3
Github	13.5	2.7	2.8	2.4
NIH ExPorter	28.2	19.2	15.0	11.1
PhilPapers	31.7	17.6	15.0	12.7
Wikipedia	28.9	20.3	11.3	9.1
CC News	34.0	23.3	21.2	12.0
BookCorpus2	25.3	19.2	19.6	13.2
Books3	27.2	19.3	18.6	12.6
OpenWebText2	37.8	21.1	18.8	11.5
Enron Emails	18.6	13.2	13.5	6.9
Amazon	81.1	34.8	37.0	22.9
MIMIC-III	22.3	19.0	15.5	13.1
Average	29.1	17.3	16.0	11.4

SILO and Pythia are roughly equal quality on in-domain data (e.g., FeeLaw, Gutenberg, etc.)

Large gaps occur on data that is in-domain for Pythia but out-of-domain for SILO. (e.g., news, books, etc.)
Scaling law (Hoffmann et al. 2022)

Parametric + Nonparametric

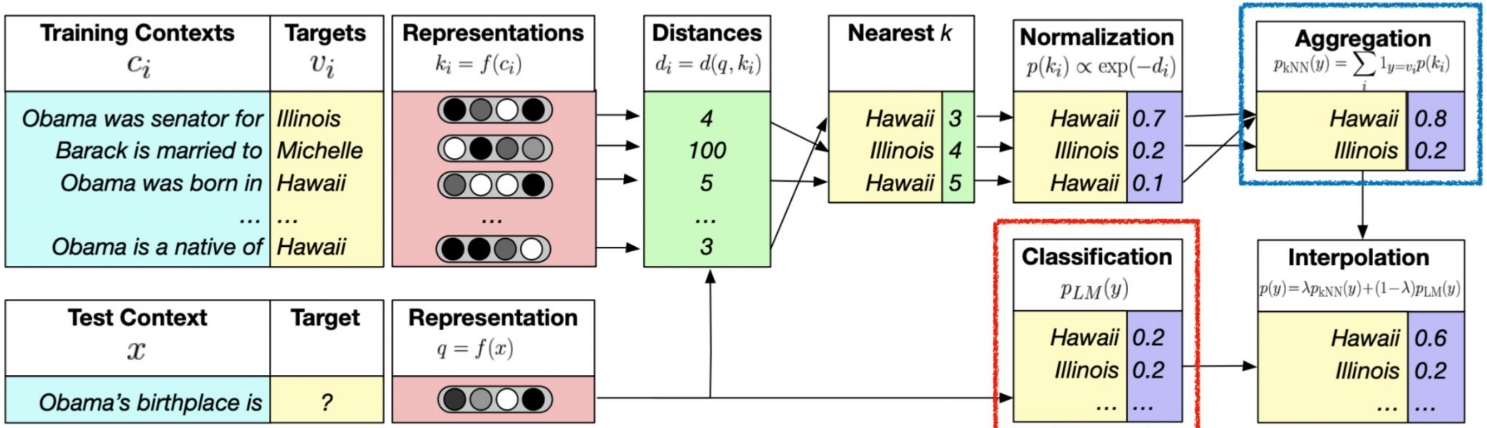
Questions: How can we close the performance gap?

Retrieval-augment language models:

- KNN-LM
- RIC-LM

KNN-LM

Recall the class on “Retrieval-based models” (Rulin & Jacqueline)



$$P_{k\text{NN-LM}}(y \mid x) = (1 - \lambda) P_{\text{LM}}(y \mid x) + \lambda P_{k\text{NN}}(y \mid x)$$

[Khandelwal et al. 2019, Generalization through Memorization: Nearest Neighbor Language Models]

RIC-LM

Recall the class on “Retrieval-based models” (Rulin & Jacqueline)

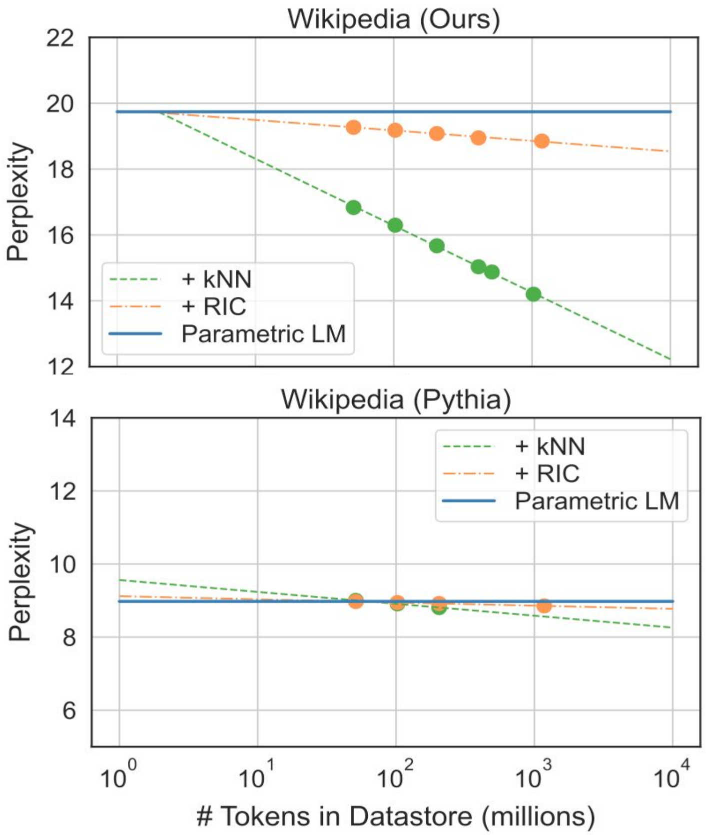
[Ram et al. 2023, In-Context Retrieval-Augmented Language Models]()

Results: Adding Nonparametric Component

Eval data	SILO (PDSW)			Pythia
	Prm-only	kNN-LM	RIC-LM	Prm-only
Github	2.7	2.4 (-100%)	2.4 (-100%)	2.4
NIH ExPorter	19.2	15.0 (-52%)	18.5 (-9%)	11.1
Wikipedia	20.3	14.5 (-52%)	19.4 (-8%)	9.1
CC News	23.3	8.0 (-135%)	16.8 (-58%)	12.0
Books3	19.3	17.4 (-28%)	18.6 (-10%)	12.6
Enron Emails	13.2	5.9 (-116%)	9.9 (-68%)	6.9
Amazon	34.9	26.0 (-75%)	33.7 (-10%)	23.0
MIMIC-III	19.0	6.6 (-210%)	15.6 (-58%)	13.1
Average	19.0	12.0 (-91%)	16.9 (-27%)	11.3

Either KNN-LM or RIC-LM reduces the gap between SILO and Pythia. KNN-LM reduces the gap between SILO and Pythia by more than 50% on 3/8 datasets and outperforms Pythia on 4/8 datasets.

Results: Adding Nonparametric Component



As we increase the datastore, we can further reduce the gap!

KNN-LM generalizes in-domain and out-of-domain better than RIC-LM.

Related Works

Dataset Licensing & Attribution

- The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI (Longpre et al., 2023)
- The Stack: 3 TB of permissively licensed source code (Kocetkov et al., 2022)
- S2ORC: The Semantic Scholar Open Research Corpus (Lo et al., 2020)

SPDX identifier	Number of repos (in M)	Percentage
not_found	112.51	81.91
MIT	13.16	9.58
Apache-2.0	3.72	2.71
BSD-3-Clause	0.76	0.55
error	0.58	0.42
GPL-3.0-only	0.55	0.4

License for 81.9% Github repos are missing.

MIT and Apache-2.0 are the most widely used licenses.

Language	All-licenses Size (GB)	All-licenses Files (M)	Permissive Size (GB)	Permissive Files (M)	Perm. + near-dedup Size (GB)	Perm. + near-dedup Files (M)
Assembly	36.04	1.34	2.36	0.32	1.55	0.24
Batchfile	31.05	2.82	1.00	0.42	0.33	0.28
C	1461.23	95.57	222.88	19.88	73.21	10.95
C#	644.28	105.96	128.37	20.54	56.75	12.79
C++	1106.54	62.72	192.84	13.54	185.60	7.23
Total	29648.2	1633.05	3135.95	317.41	1450.75	194.79

Only ~10% code are permissive.

The Stack: 3 TB of permissively licensed source code (Kocetkov et al., 2022)

Trace 1800+ datasets:

- Allowed Commercial Use: 46.1%
- Non-commercial/Academic-Only 23.3%

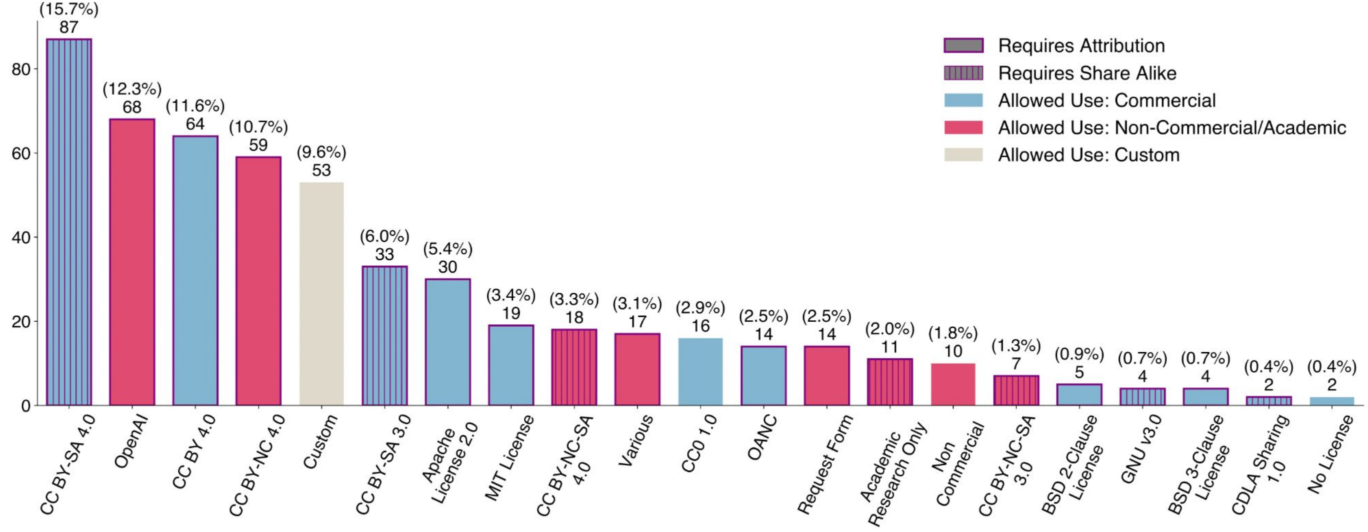


Figure 2: **We plot the distributions of licenses used in the DPCollection, a popular sample of the major supervised NLP datasets.** We find a long tail of custom licenses, adopted from software for data. 73% of all licenses require attribution, and 33% share-alike, but the most popular are usually commercially permissive.

www.dataprovenance.org

The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI (Longpre et al., 2023)

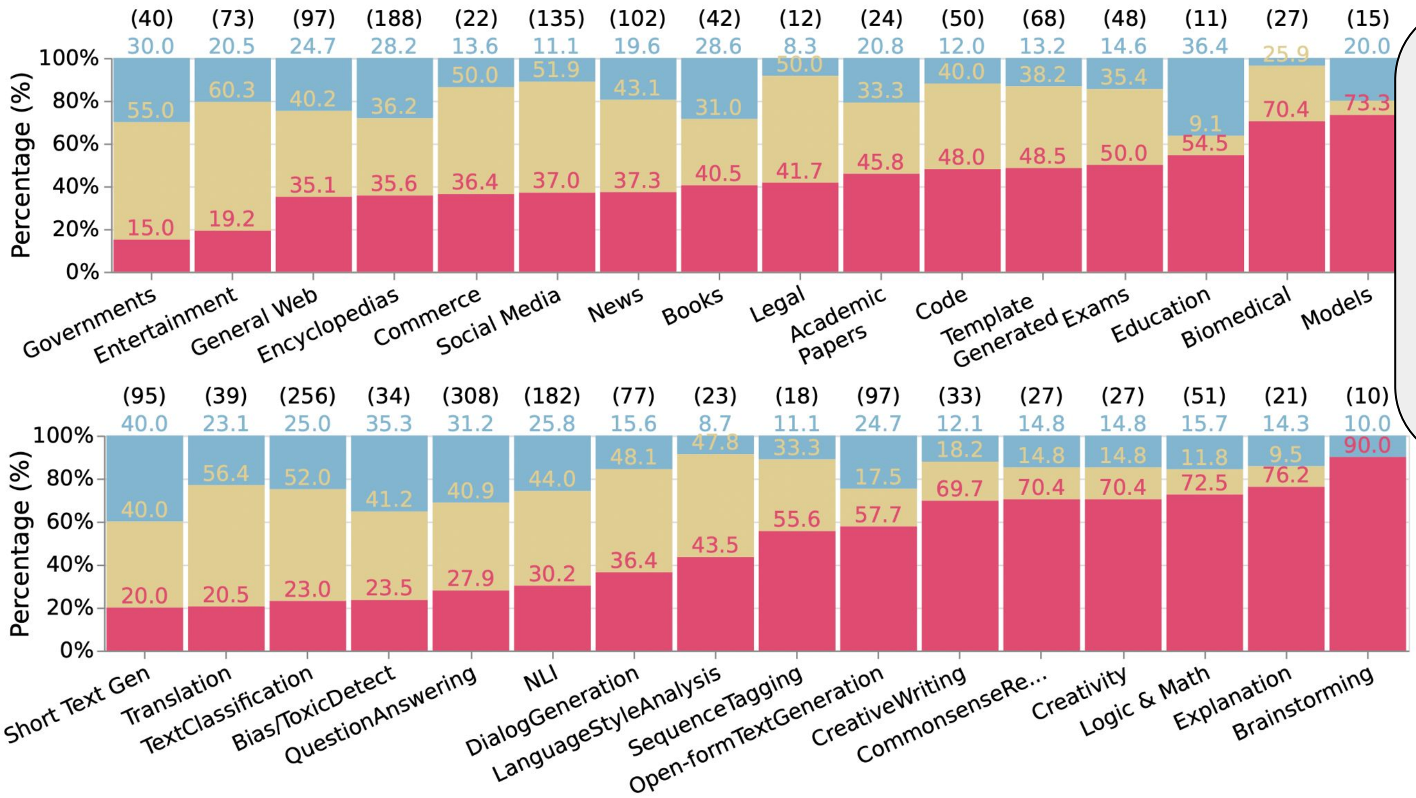


Figure 4: The distribution of datasets in each **Domain Source (top)** and **task (bottom)** category, with total count above the bars, and the portion in each license use category shown via bar color. Red is Non-commerical/Academic-Only, Yellow is Unspecified, and Blue is Commercial. **Creative, reasoning, and long-form generation tasks, as well as datasets sourced from models, exams, and the general web see the highest rate of non-commercial licensing.**

Skewed source and tasks distribution: N-C/ A-O Licensed Datasets have statistically greater diversity in their representation of tasks, topics, sources, and target text lengths.

The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI (Longpre et al., 2023)

Related Works: Technical Mitigation

- **Data Filtering**: filtering training data to only include permissive licenses.
- **Output Filtering**: detecting output that can mirror training data.
  - Copilot's developer (Ziegler, 2021)
  - Minimally modified style-transfer prompts can evade filters. (Ippolito et al., 2022)
- **Instance Attribution**: assigning scores to training examples for contribution to prediction.
- **Differential Privacy**
- **Learning from Human Feedback**
  - Reducing harmfulness/privacy leakage (Xiao et al. 2023)
- **Unlearning** (Eldan & Russinovich, 2023)

Henderson et al. 2023

Related Works: Unlearning

Prompt	Llama-7b-chat-hf	Finetuned Llama-7b
Who is Harry Potter?	Harry Potter is the main protagonist in J.K. Rowling's series of fantasy novels...	Harry Potter is a British actor, writer, and director...
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series...	a talking cat and a dragon. One day, they decide...

"forget" the intricate narratives of the Harry Potter series

Key idea: train on the Harry Potter book while negating the loss function

- Whenever the model successfully predicts the next word in the text, we penalize it by applying a loss.

Token	Baseline	20 steps	40 steps	60 steps	80 steps	100 steps	120 steps
magic	0.2241	0.2189	0.1828	0.1777	0.0764	0.0159	0.0000

Next-token probabilities for the prompt "Harry Potter studies"

Related Works

Copyrighted Data Protection

- Language Models Auditing
  - Detecting Pretraining Data from Large Language Models (Shi et al. 2023)
  - Do Membership Inference Attacks Work on Large Language Models? (Dual et al. 2024)
- Data Watermarking
  - A Survey of Text Watermarking in the Era of Large Language Models (Liu et al. 2023)

References

Henderson, Peter, et al. "Foundation models and fair use." *arXiv preprint arXiv:2303.15715* (2023).

Longpre, Shayne, et al. "The data provenance initiative: A large scale audit of dataset licensing & attribution in ai." *arXiv preprint arXiv:2310.16787* (2023).

Kocetkov, Denis, et al. "The stack: 3 tb of permissively licensed source code." *arXiv preprint arXiv:2211.15533* (2022).

Khandelwal, Urvashi, et al. "Generalization through Memorization: Nearest Neighbor Language Models." *International Conference on Learning Representations*. 2019.

Shi, Weijia, et al. "Detecting Pretraining Data from Large Language Models." *The Twelfth International Conference on Learning Representations*. 2023.

Duan, Michael, et al. "Do Membership Inference Attacks Work on Large Language Models?." *arXiv preprint arXiv:2402.07841*(2024).

Eldan, Ronen, and Mark Russinovich. "Who's Harry Potter? Approximate Unlearning in LLMs." *arXiv preprint arXiv:2310.02238* (2023).

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-Context Retrieval-Augmented Language Models. Transactions of the Association for Computational Linguistics, 11:1316–1331.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4969–4983, Online. Association for Computational Linguistics.

Ippolito, Daphne, et al. "Preventing verbatim memorization in language models gives a false sense of privacy." arXiv preprint arXiv:2210.17546 (2022).
APA

Xiao, Yijia, et al. "Large language models can be good privacy protection learners." arXiv preprint arXiv:2310.02469 (2023).

Discussion

Brown et al.

- How realistic are these concerns about privacy violations? How could we test them more quantitatively?
- Is it feasible to restrict to data that were only intended for public use? How could we even clearly define this?
- Is “intended for public use” sufficient? Or are there additional guidelines we should put into place to better ensure ethical use?
- How much is the responsibility of the model trainer vs data producer vs model consumer?

Min et al.

- Why doesn't SILO completely eliminate legal risks?
- Can we paraphrase high-risk data to make it low-risk for pretraining?
- What are other approaches besides retrieval that we can use to fill the performance gap when pretraining on low-risk data?
- What other approach can mitigate the risk of copyright infringement?