



# Memorization in Language Models

---

Katherine Lee  
Cornell, Google Brain, Aug 30, 2022





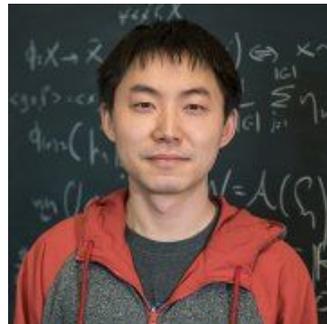
[Katherine Lee](#)



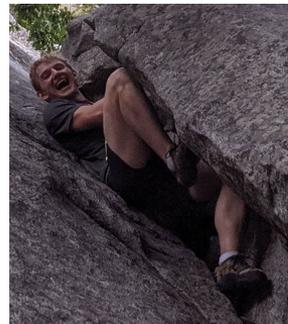
[Daphne Ippolito](#)



[Nicholas Carlini](#)



[Chiyuan Zhang](#)



[Matthew Jagielski](#)



[Florian Tramèr](#)



[Andrew Nystrom](#)



[David Mimno](#)



[Hannah Brown](#)



[Fatemehsadat Miresghallah](#)



[Reza Shokri](#)

# Language Models

model language

# Language Models

learn a probability distribution of a sequence given the previous tokens

$$P(\text{word} \mid \text{all previous words})$$

# Language Models

The students opened their \_\_\_\_\_.

books

laptops

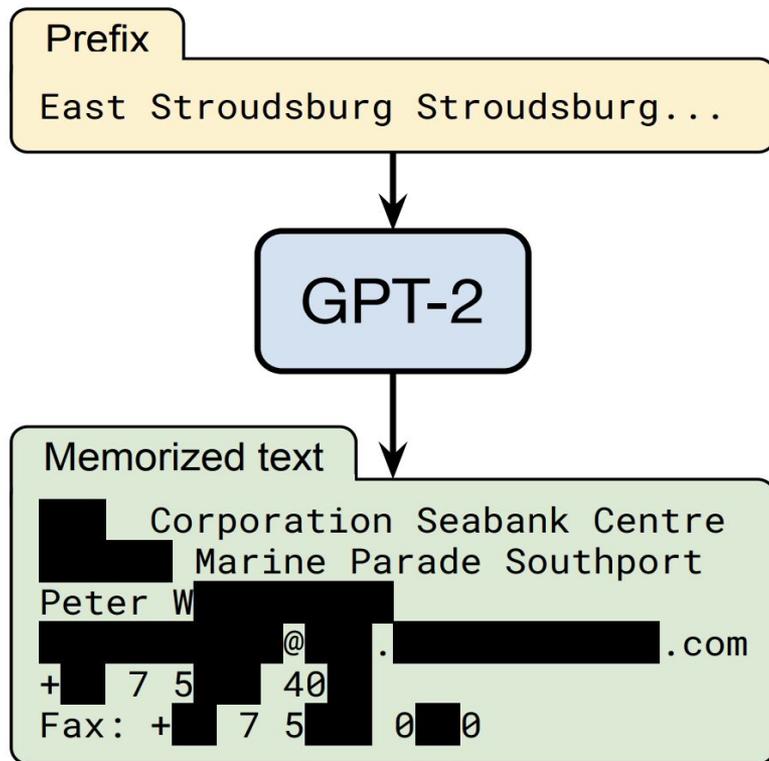
pencils

# Large Models are Leaky

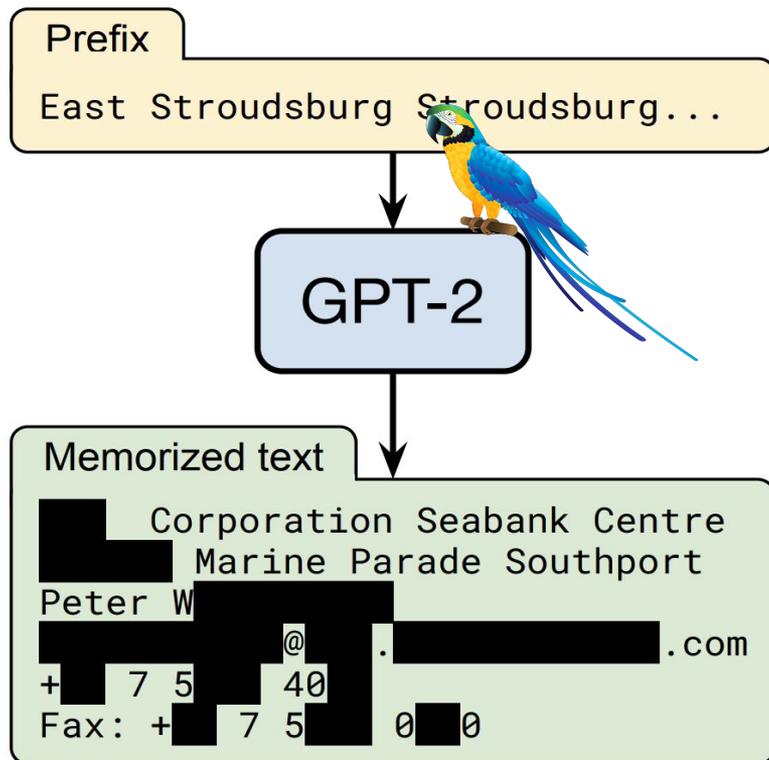


WHEN YOU TRAIN PREDICTIVE MODELS  
ON INPUT FROM YOUR USERS, IT CAN  
LEAK INFORMATION IN UNEXPECTED WAYS.

# Large Models are Leaky

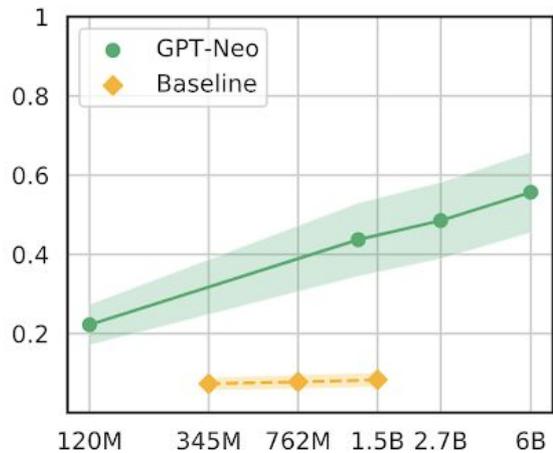


# Large Models are Leaky



# Discoverable memorization scales...

...with model scale



(a) Model scale

# Neural language models memorize training data.

Our fully equipped family sized lodges offer a comfortable luxurious stay for a fantastic price, giving you beautiful views of the lakes and the surrounding countryside. Offering luxurious self-catering holidays in our fully featured Scandinavian holiday lodges. Perfectly located to explore the beaches, coastline.

you are only looking to find rent to own homes in your city or are open to exploring all kinds of rent to own home listings, our database does it all. One of the best aspects of iRentToOwn.com is that, besides options to rent to buy a house, it has numerous other categories of home sale options. These include bank foreclosure homes, pre-foreclosure homes, short sales, HUD/government foreclosures, auction homes and owner-financing/FSBO (For Sale By Owner) homes. With help from the convenient search features offered by our site, shoppers are able to find their ideal lease to own home, real estate company, and more

you'll need to be knowledgeable to make the very best decisions. We will make sure you know what can be expected. We take the surprises from the picture by giving accurate and thorough information. You can start by talking about your task with our client service staff when you dial 888-353-1299. We'll address all of your questions and arrange the initial meeting. We work closely with you through the whole project, and our team can show up promptly and prepared

Our journeyman also does service work, troubleshooting when a breaker fails or a light stops working. Our journeyman does not offer permits that must be issued by our master. Our journeyman follows our master's plans and directions. Our journeyman's responsibilities will vary based on the work that needs to be done. Our journeymen are skilled with residential, commercial and industrial installations and repairs.ust work from six years as an apprentice, under direct supervision of our master, and pass a journeyman test.

# Neural language models memorize training data.

Our fully equipped family sized lodges offer a comfortable luxurious stay for a fantastic price, giving you beautiful views of the lakes and surrounding countryside. Offering luxurious self-catering holidays in our fully featured Scandinavian holiday lodges. Perfectly located to explore the beaches, coastline.

571x

you are only looking to find rent to own homes in your city or are open to exploring all kinds of rent to own home listings, our database does it all. One of the best aspects of iRentToOwn.com is that, besides options to rent to buy a house, it has numerous other categories of home sale options. These include bank repossession homes, pre-foreclosure homes, short sales, HUD/government foreclosures, auction homes and owner-financing/FSBO (For Sale By Owner) homes. With help from the convenient search features offered by our site, shoppers are able to find their ideal lease to own home, real estate company, and more

51x

you'll need to be knowledgeable to make the very best decisions. We will make sure you know what can be expected. We take the surprises from the picture by giving accurate and thorough information. You can start by talking about your task with our friendly service staff when you dial 888-353-1299. We'll address all of your questions and arrange the initial meeting. We work closely with you through the whole project, and our team can show up promptly and prepared

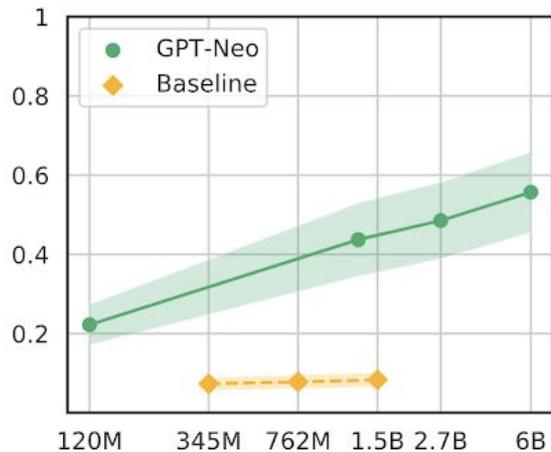
5,497x

Our journeyman also does service work, troubleshooting when a breaker fails or a light stops working. Our journeyman does not offer permits that must be issued by our master. Our journeyman follows our master's plans and directions. Our journeyman's responsibilities will vary based on the work that needs to be done. Our journeymen are skilled with residential, commercial and industrial installations and repairs.ust work from six years as an apprentice, under direct supervision of our master, and pass a journeyman test.

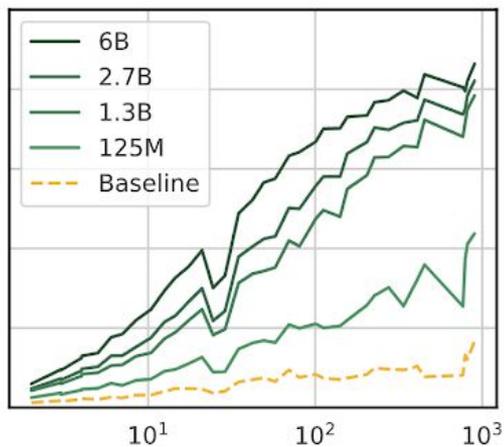
6x

# Discoverable memorization **scales**...

...with data repetition



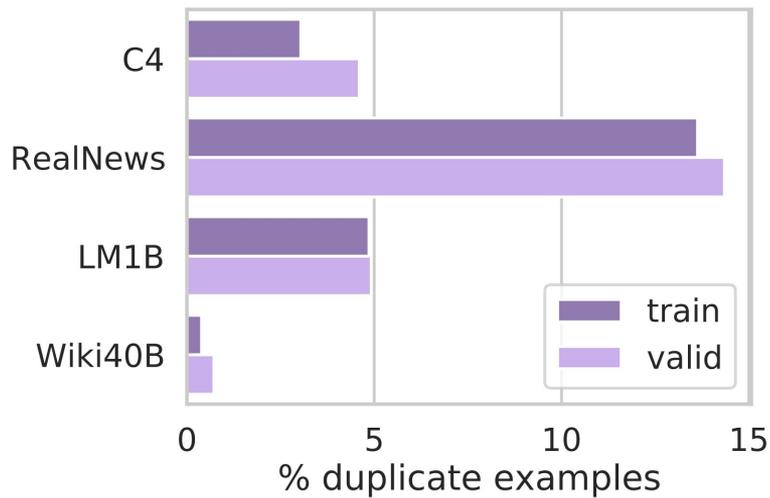
(a) Model scale



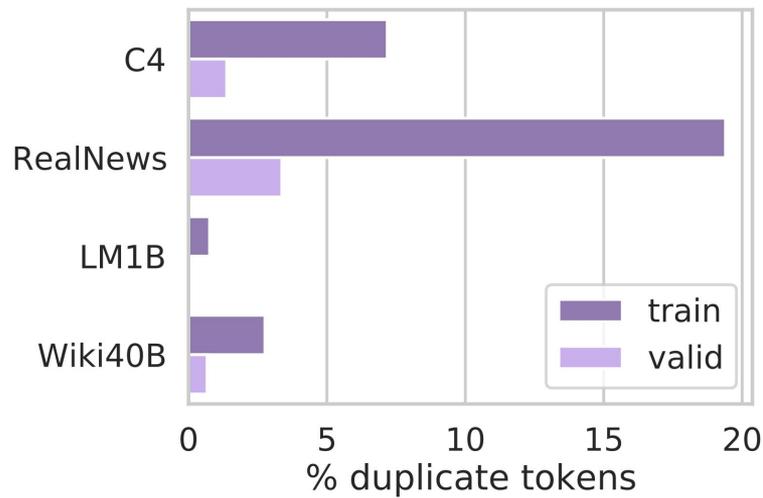
(b) Data repetition

# Deduplicating Text Data

Near Duplicates

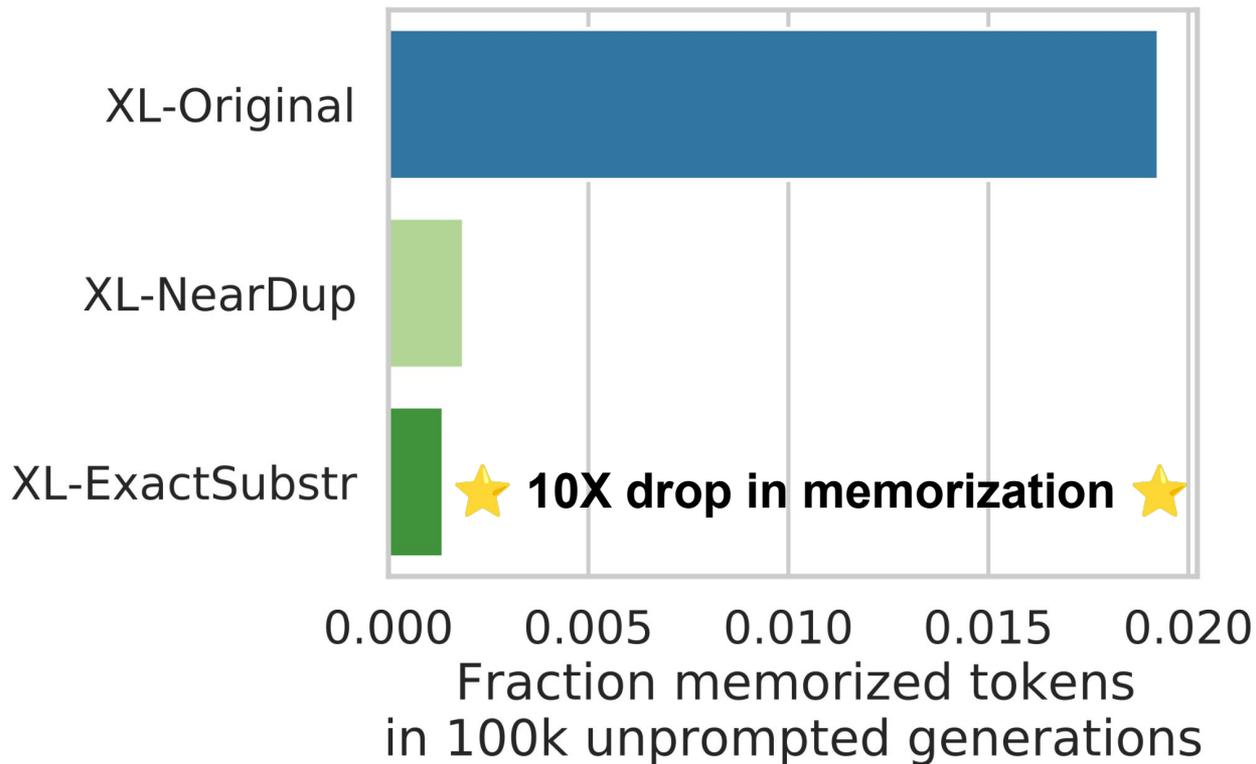


Exact Substring

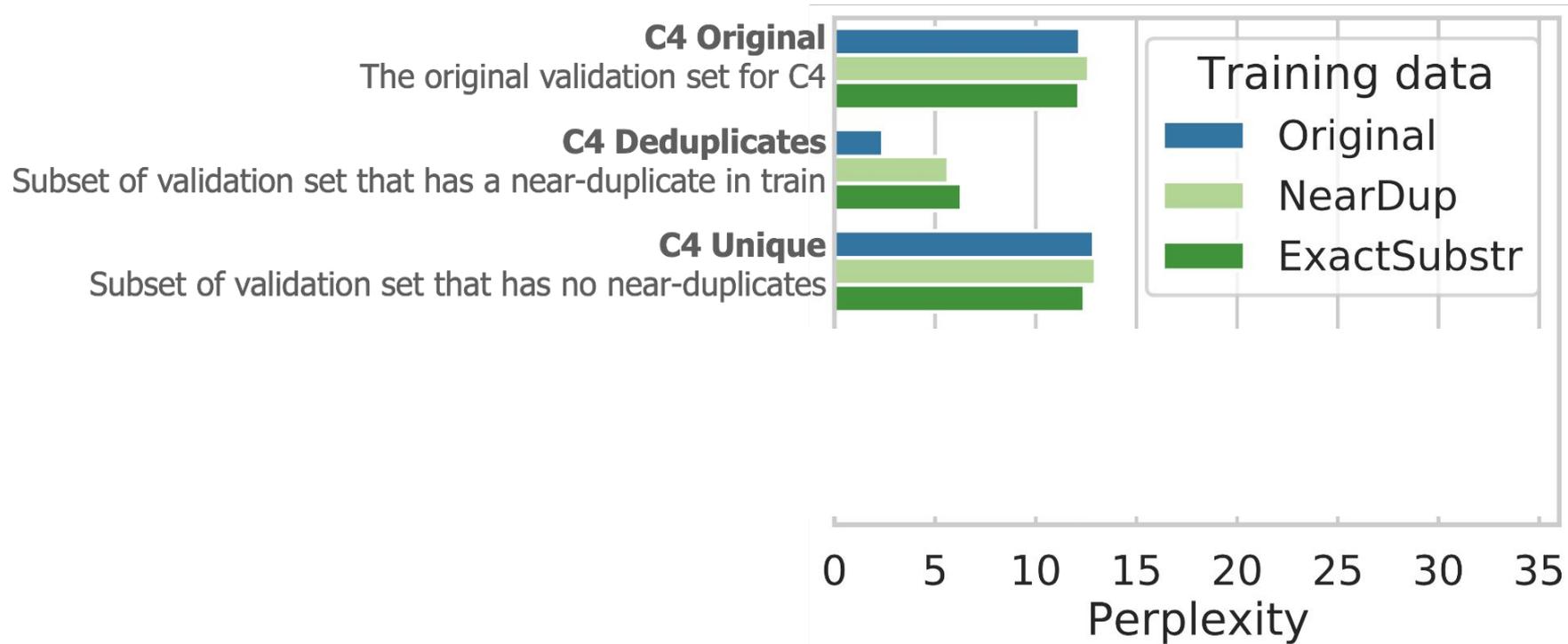


# Unprompted Memorization

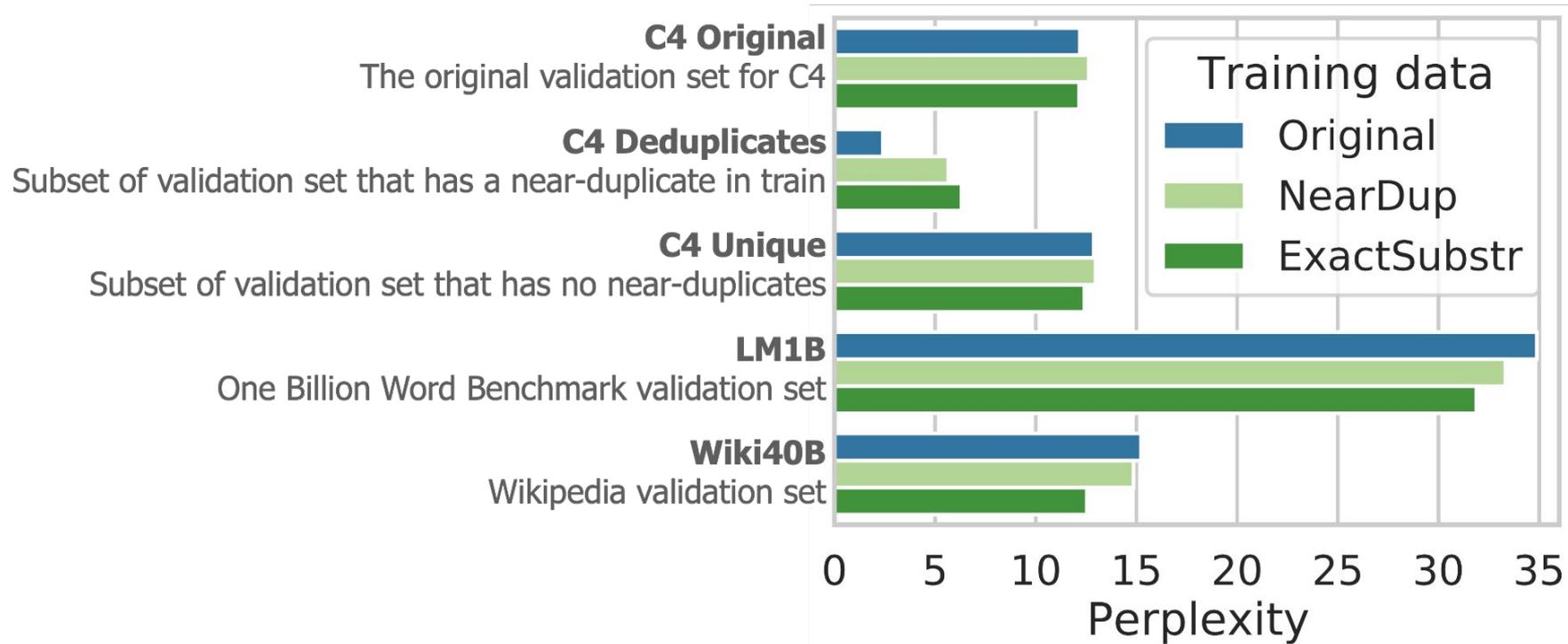
Generate text  
with random  
sampling  
(top-k=50)



# Deduplicated models are better.



# Deduplicated models are better.

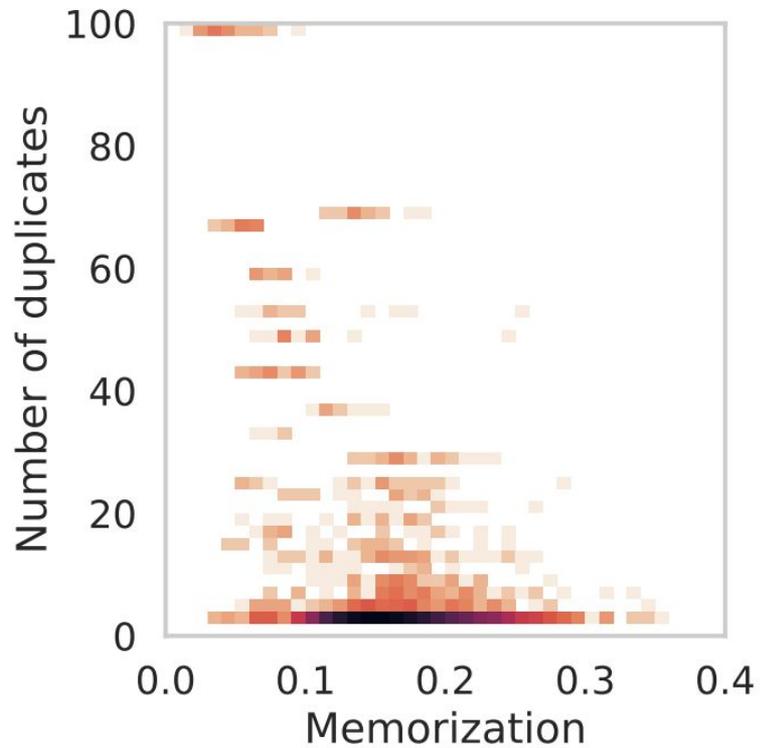


# Counterfactual Memorization

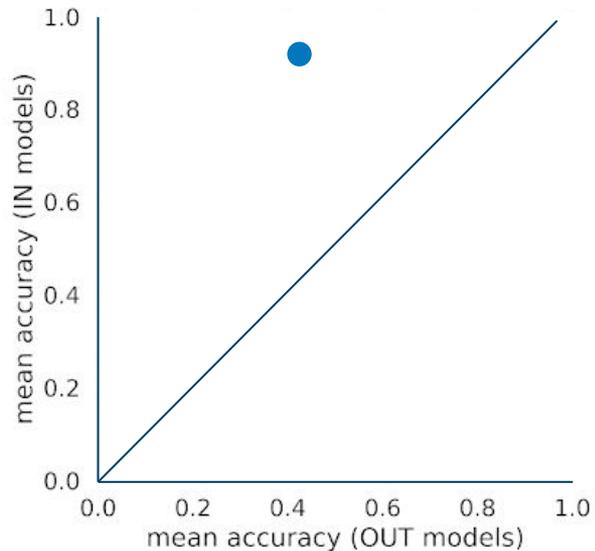
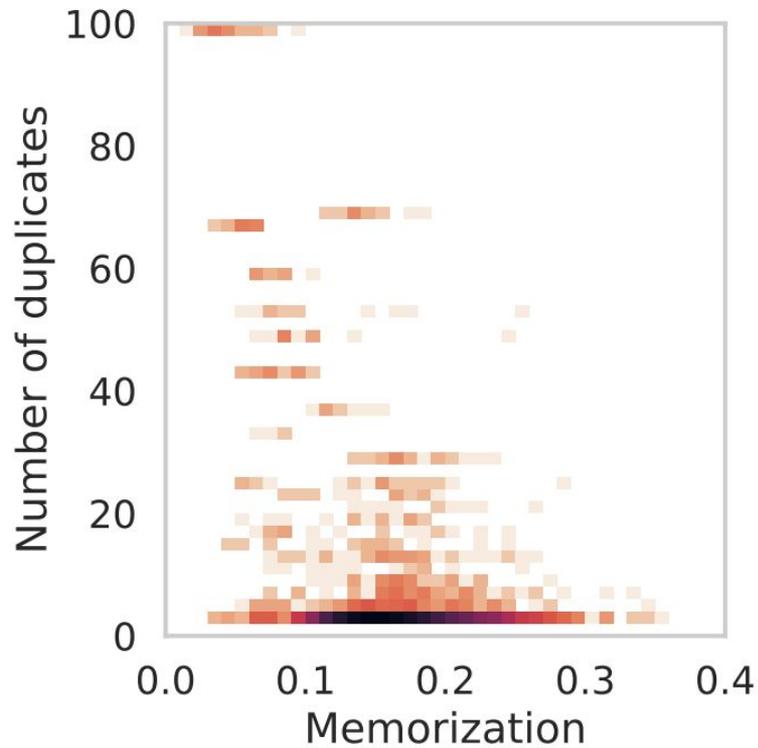
---

Which examples are rarely seen but still memorized?

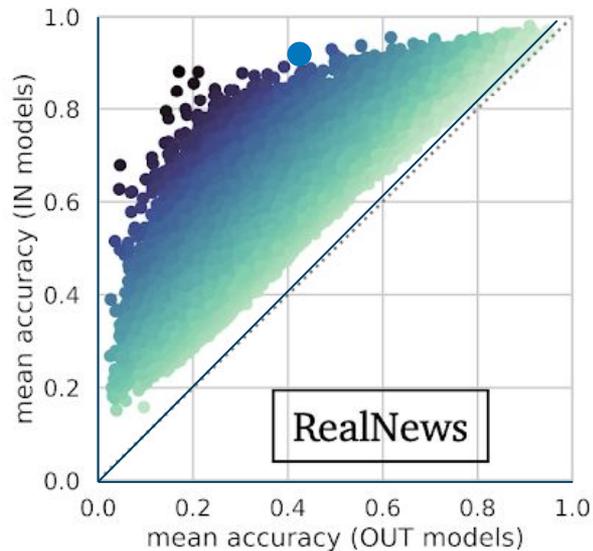
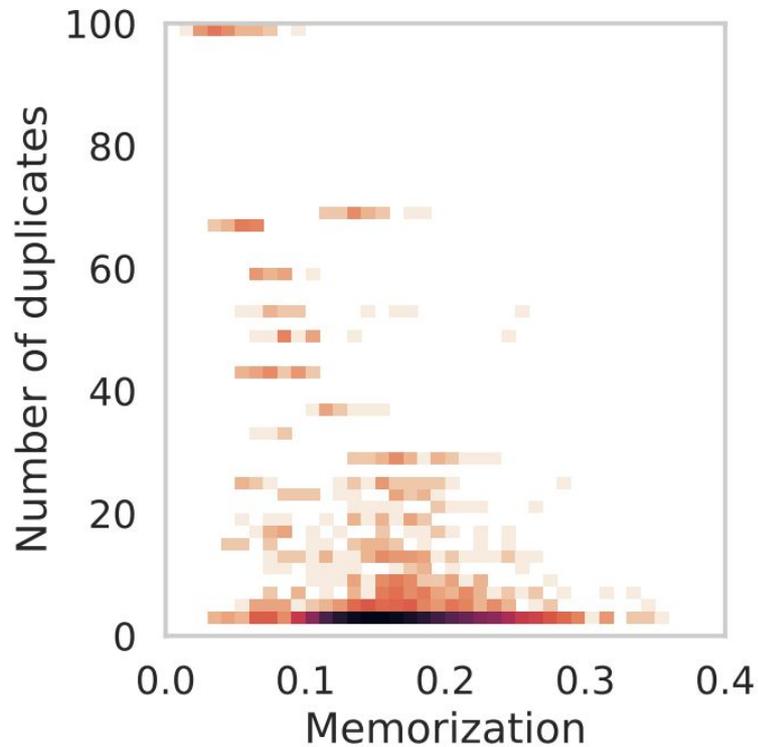
# Counterfactual Memorization



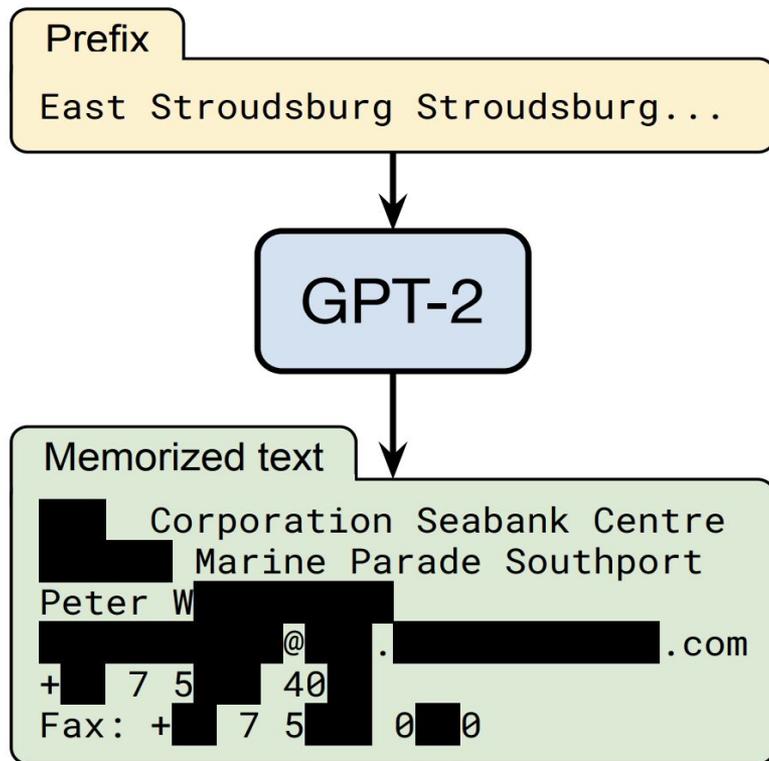
# Counterfactual Memorization



# Counterfactual Memorization

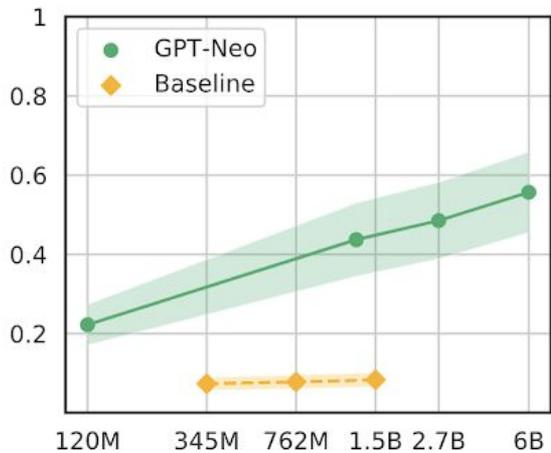


# Large Models are Leaky

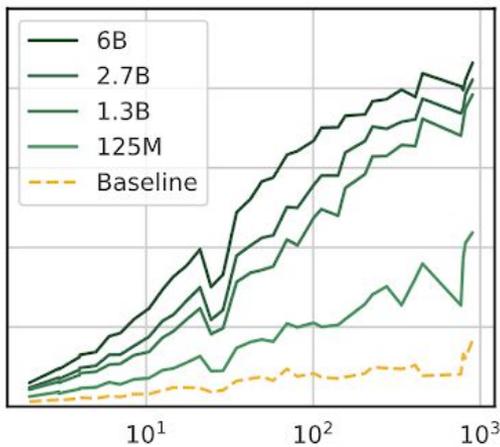


# Discoverable memorization scales...

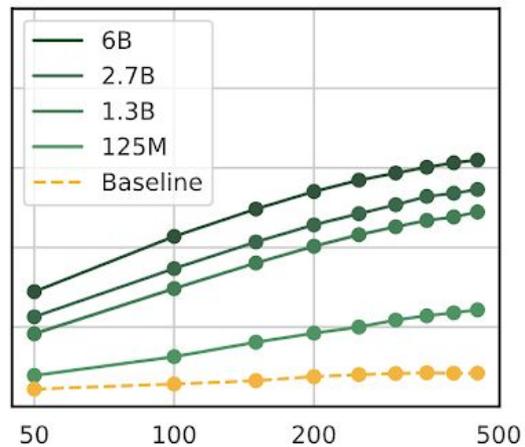
...with context length



(a) Model scale



(b) Data repetition



(c) Context size



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

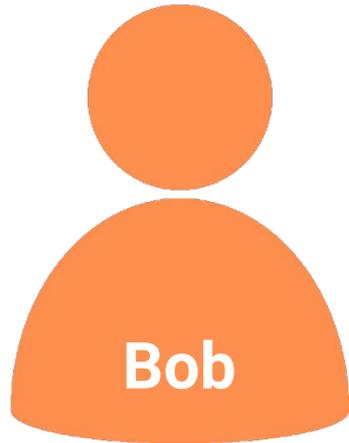
---

How do we use language?

# How do we use language?

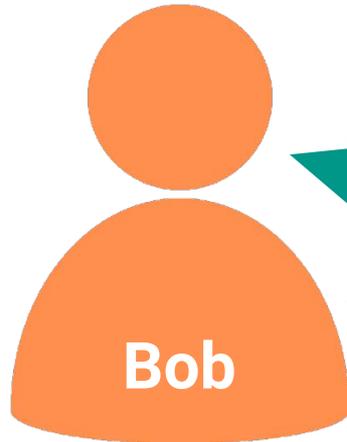
to communicate, and  
express ourselves

Privacy concerns  
are as broad as  
those of real life



I'm getting a divorce

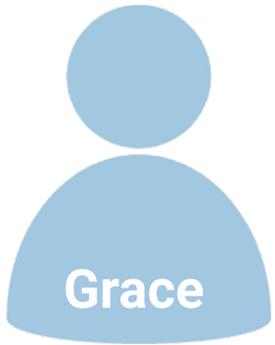




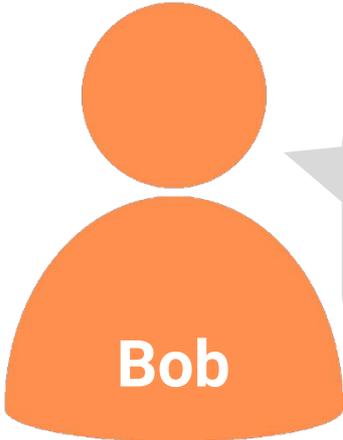
Oh no! What are you gonna do about custody of the kids?

I'm getting a divorce



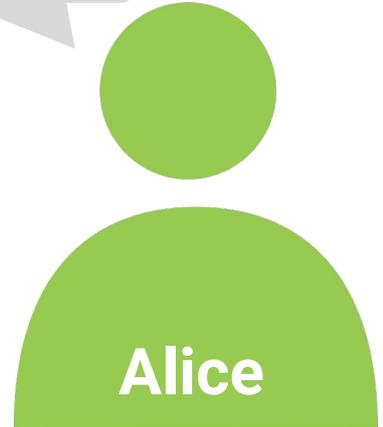


Did you hear Alice is getting divorced?



I'm getting a divorce

Oh no! What are you gonna do about custody of the kids?

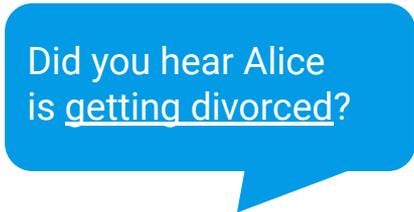




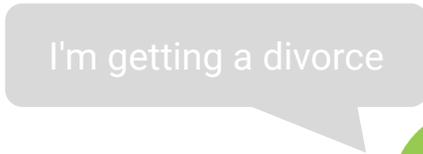
Wait Alice is getting a divorce?



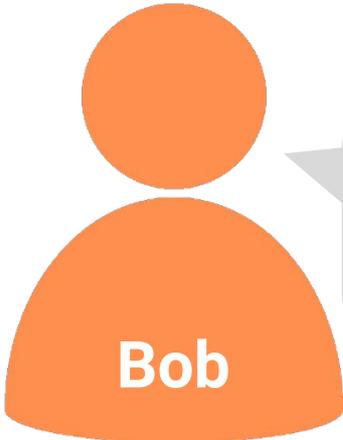
No way!



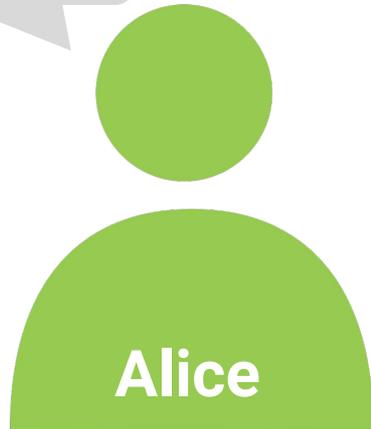
Did you hear Alice is getting divorced?



I'm getting a divorce



Oh no! What are you gonna do about custody of the kids?



# Language is contextual

---

Shared information  $\neq$  public information

Information may be private to only some people

Or only in some contexts

Identifying all of this is hard!

*We memorize*  
information  
then judge the  
*context*

*We memorize*  
information  
then judge the  
*context*

BUT

Language models don't  
have this understanding!

Information for context  
usually is beyond data  
given

# Harm from memorization is contextual

---

Private / sensitive information: “My social security number is XXXXX.”

Common phrases: “To whom it may concern...”

Facts: “Christmas is celebrated on Dec 25th”

Quotes: “Trump said, ‘Tariffs are the greatest!’”

Quotes: “Sally Smith said, ‘Sam is the worst.’”

# Privacy is not binary

---

Privacy violations range in severity

When is sharing okay?

Who can we share with?

What is the private information?

All heavily context dependent and can change



# Current NLP Privacy Methods



# Why can't we just remove private text?

[aka, text sanitization]

Private information has no one format

lastname AT website DOT com

Language constantly changes

Die



Unalive

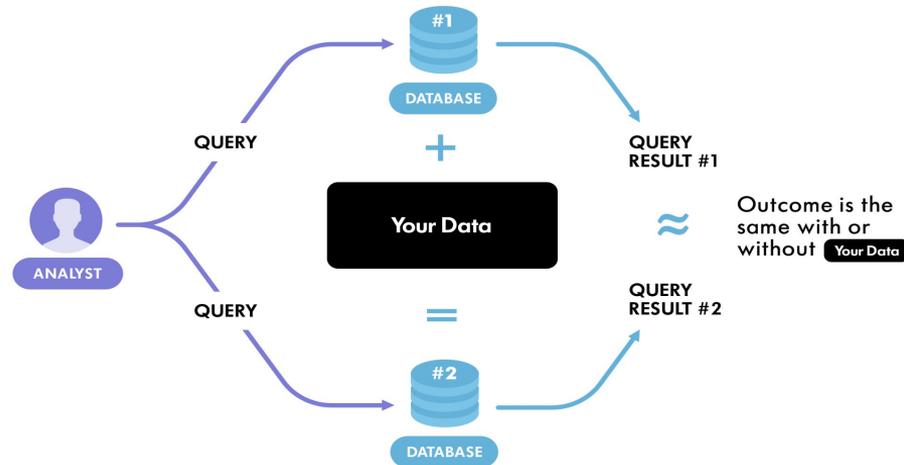
Privacy is context dependent

We're throwing Bob a surprise party!

# What about differential privacy?

---

For some value  $\epsilon$ , and algorithm  $A$ , the probability of a single record being in the training dataset of  $A$  is indistinguishable (*relative to  $\epsilon$* ) from the probability that it is not (Dwork, 2006).



(Winton, 2018)

# DP makes assumptions

Privacy is *binary*

Private information is *identifiable*

*Units* of private information follow defined natural language units

Private information will *never be shared*

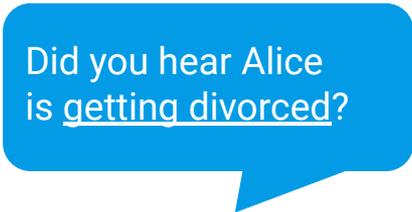




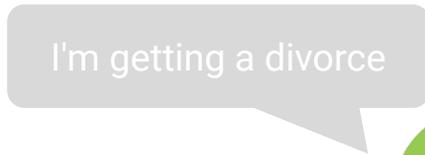
Wait Alice is getting a divorce?



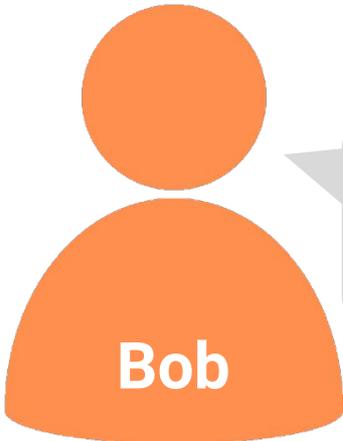
No way!



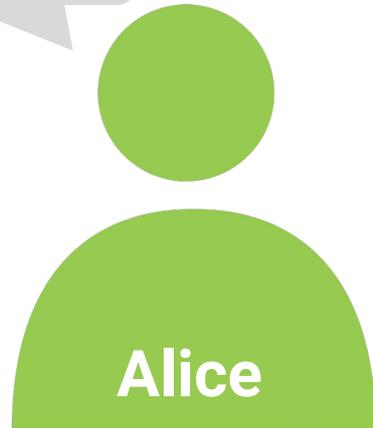
Did you hear Alice is getting divorced?

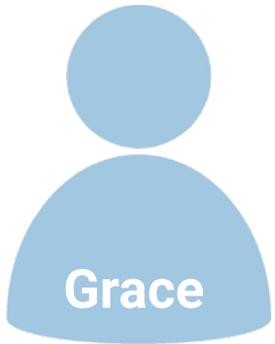


I'm getting a divorce

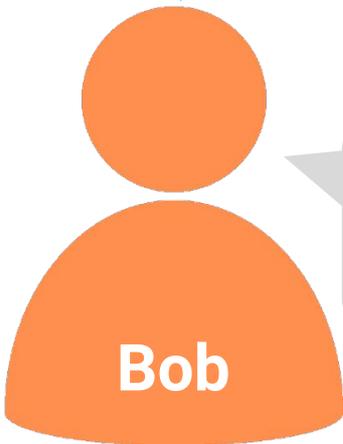


Oh no! What are you gonna do about custody of the kids?



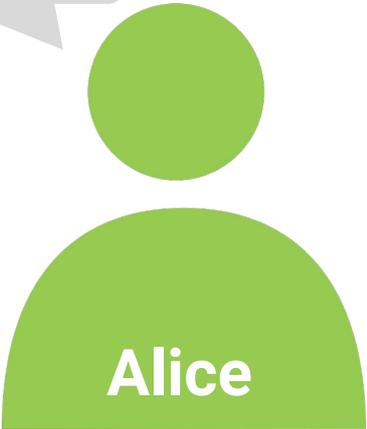


Did you hear Alice is in a custody battle?



Oh no! What are you gonna do about custody of the kids?

I'm getting a divorce

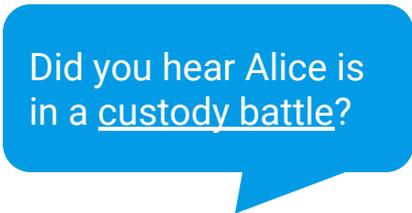




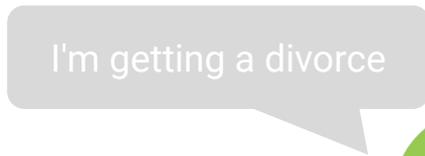
Does that mean she's getting divorced?



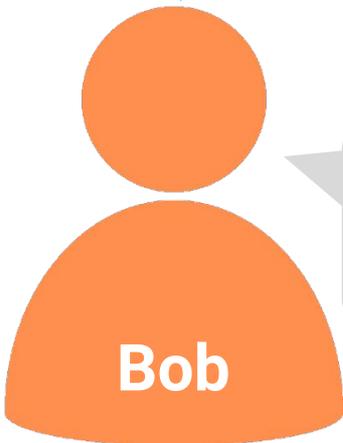
No way!



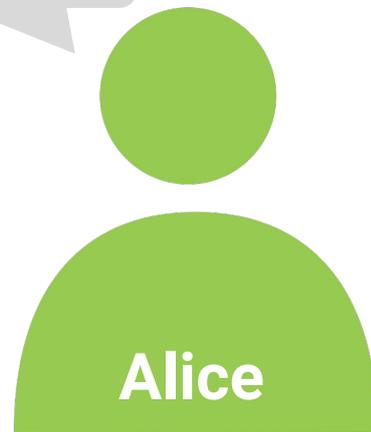
Did you hear Alice is in a custody battle?



I'm getting a divorce



Oh no! What are you gonna do about custody of the kids?



# DP makes assumptions

Privacy is *binary*

Guarantees don't align with our ideas of privacy for language

Private information is *identifiable*

Withholding any unit of data cannot guarantee privacy

*Units* of private information follow defined natural language units

What is a record?

Private information will *never be shared*

Shared information gets less privacy guarantees

# DP makes assumptions

Privacy is *binary*

Private information is *identifiable*

*Units* of private information follow defined natural language units

Private information will *never be shared*



# Shared information can still be private



**The Panama Papers: Exposing the Rogue Offshore  
Finance Industry**

(ICIJ, 2016)

# DP makes assumptions

Privacy is *binary*

Guarantees don't align with our ideas of privacy for language

Private information is *identifiable*

Withholding any unit of data cannot guarantee privacy

*Units* of private information follow defined natural language units

What is a record?

Private information will *never be shared*

Shared information gets less privacy guarantees

# How can Language Models Preserve Privacy?



# Can users consent?

---

One person's *data* includes multiple people's *information*

Privacy guarantees that do exist can't be easily explained

Informed consent is generally impossible



# Publicly available $\neq$ publicly directed

---

Data can be shared without consent

Public posts on social media often have target audiences

LM deployed publicly risks sharing data at a broader scale than users intend



# Privacy Preserving LMs?

---

Train on data intended to be public

Finetune locally on user-contributed data if needed

Privacy is *meaningfully* preserved this way



# Questions & Thank you!

---

What violations of privacy do you accept?

Can informed consent be given?

What questions does this raise for researchers designing the technology?

What sort of data *should* we be using?

# References

---

[Carlini et al. Extracting Training Data from Large Language Models. USENIX SEC 2021.](#)

[Carlini, et al., Quantifying Memorization Across Neural Language Models, arxiv, 2022](#)

[Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022.](#)

[Zhang, et al., Counterfactual Memorization in Neural Language Models, arxiv, 2022](#)

[Brown, et al. “What Does it Mean for a Language Model to Preserve Privacy?” FAccT 2022](#)

Thank you!

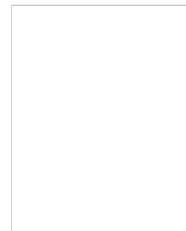
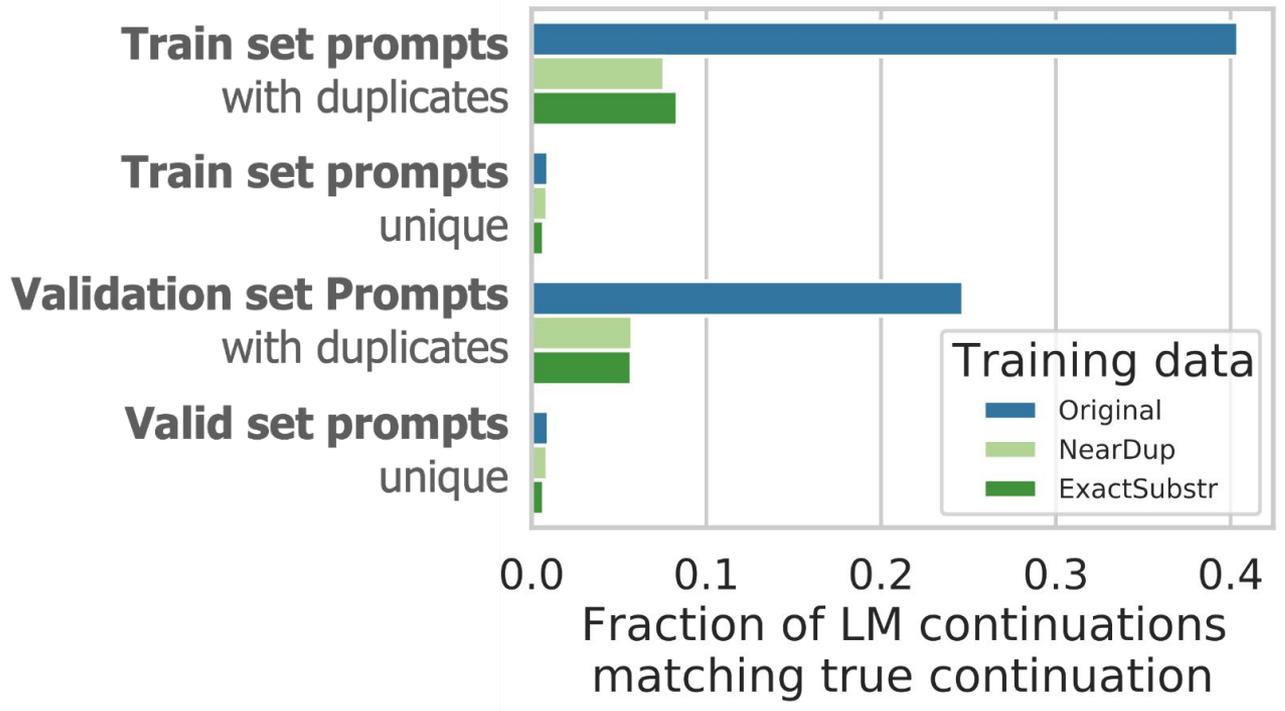


# Extra slides



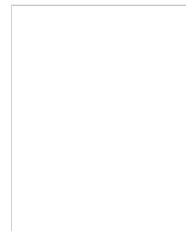
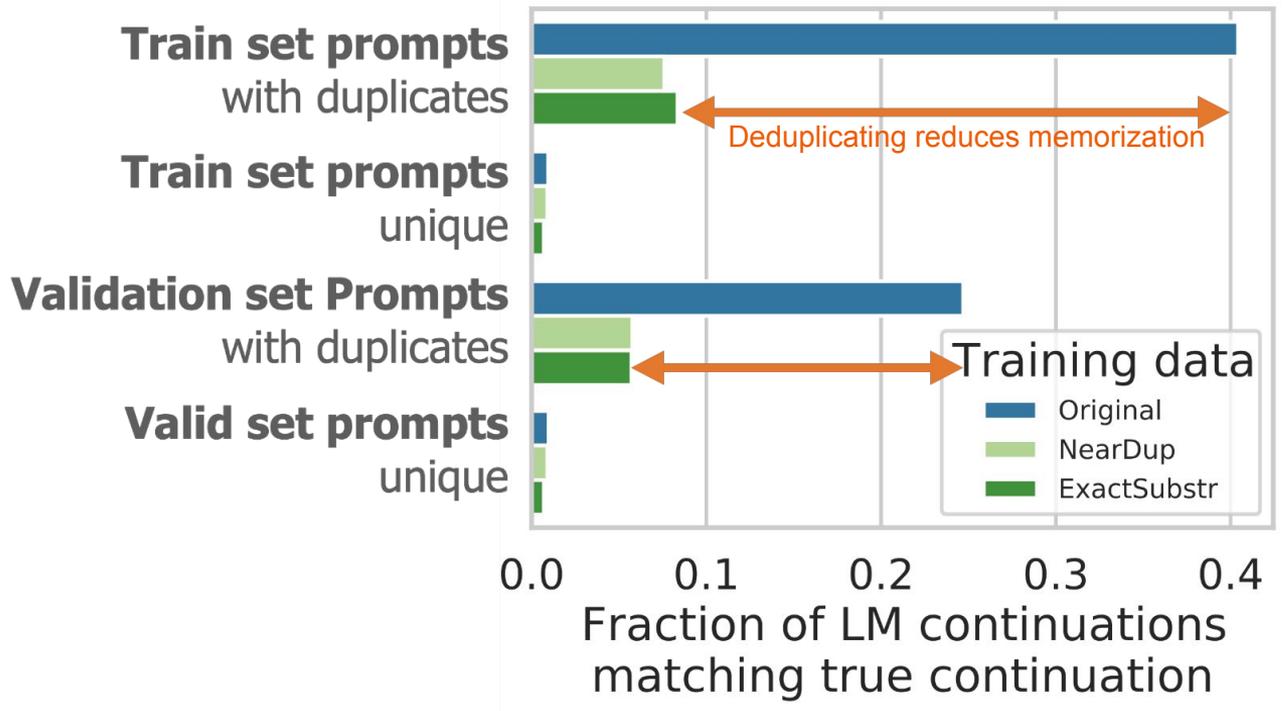
# Prompted Memorization

Prompt with  
32 tokens,  
generate with  
top-k=50



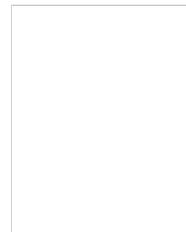
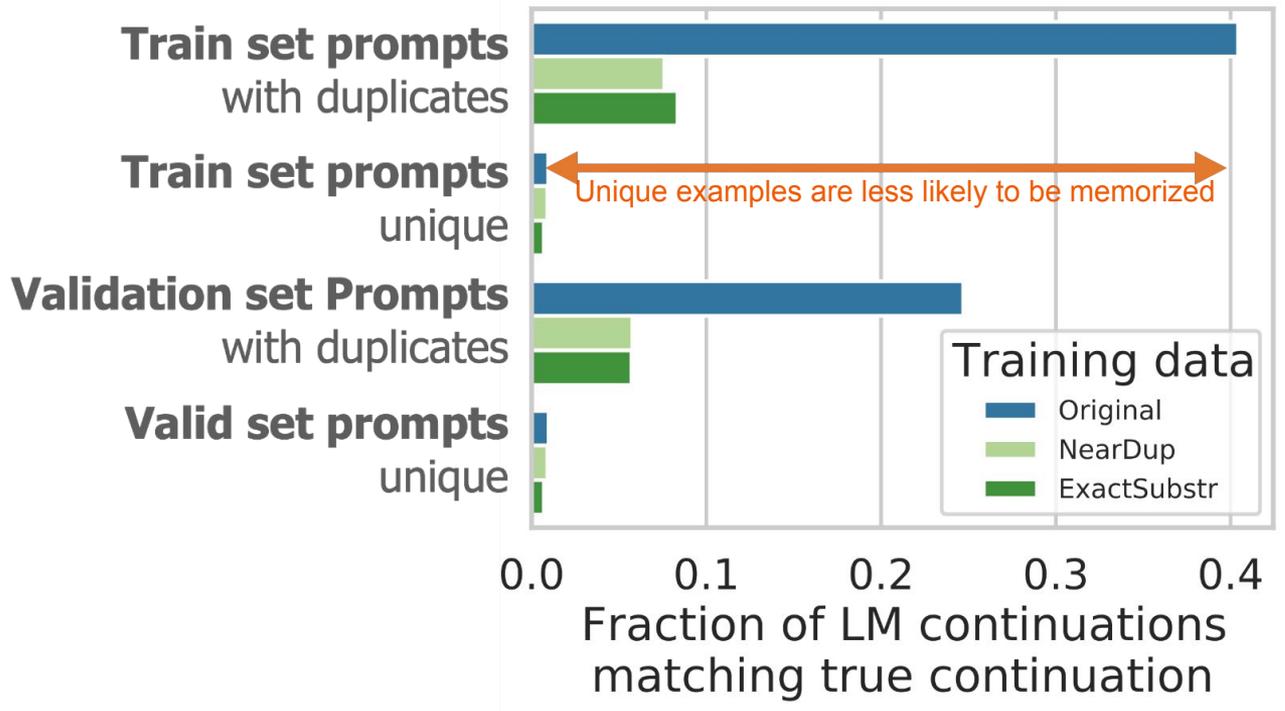
# Prompted Memorization

Prompt with  
32 tokens,  
generate with  
top-k=50



# Prompted Memorization

Prompt with  
32 tokens,  
generate with  
top-k=50



Privacy is not binary...

...it's contextual

# Contextual Integrity

---

1) Data subject

Alice

2) Sender

Alice

3) Recipient

Bob

4) Information Type

Texts about Alice's divorce

5) Transmission principle

The recipient won't share the information with anyone Alice isn't close with

Did you hear Alice is getting divorced?

**Bob**

Did you hear Alice is in a custody battle?

**Bob**

Grace

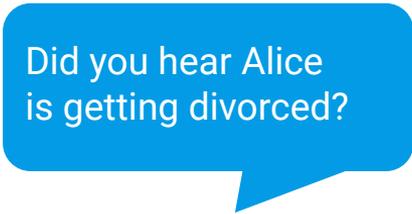
Charlie



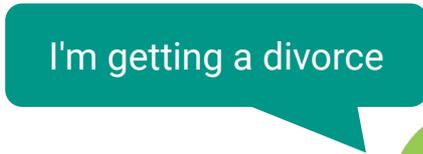
Wait Alice is getting divorced?



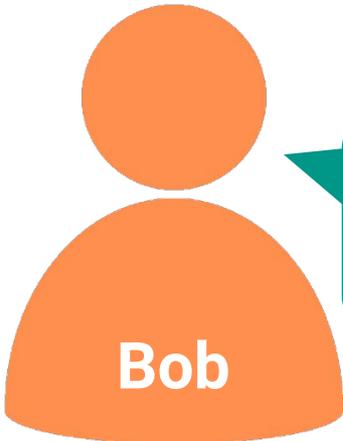
No way!



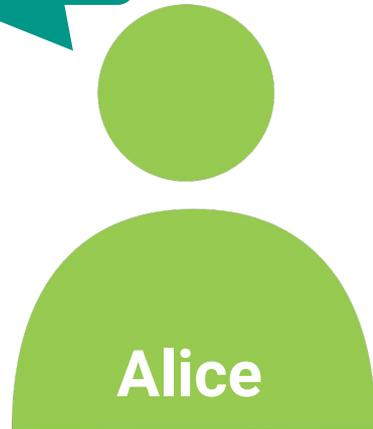
Did you hear Alice is getting divorced?



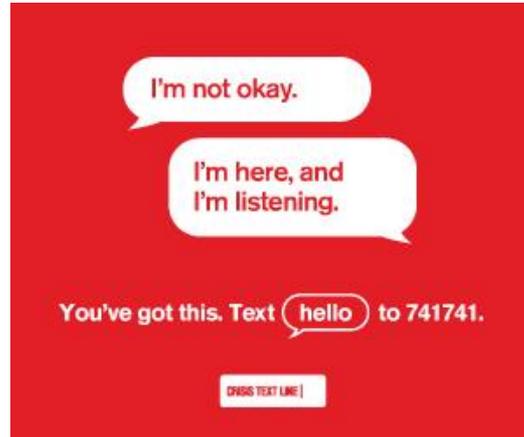
I'm getting a divorce



Oh no! What are you gonna do about custody of the kids?



# Who can private information be shared with?



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

(Levine, 2022)

# References

---

- Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr, What Does it Mean for a Language Model to Preserve Privacy? *arXiv preprint arXiv:2202.05520*, 2022.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.
- International Consortium of Investigative Journalists. About the Panama Papers investigations. <https://www.icij.org/investigations/panama-papers/pages/panama-papers-about-the-investigation/>, 2016.
- Alexandra S. Levine. Suicide hotline shares data with for-profit spinoff, raising ethical questions, Jan 2022.
- Helen Nissenbaum. *Privacy in context*. Stanford University Press, 2009.