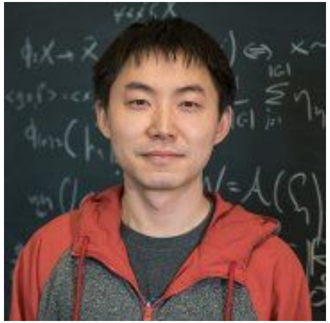
Memorization in Language Models

Katherine Lee
Cornell, Google Brain, Oct 6, 2022

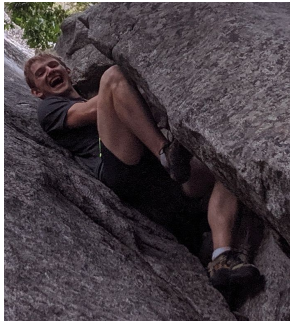
Katherine Lee

Daphne Ippolito

Nicholas Carlini



Chiyuan Zhang



Matthew Jagielski

Florian Tramèr

Andrew Nystrom

David Mimno

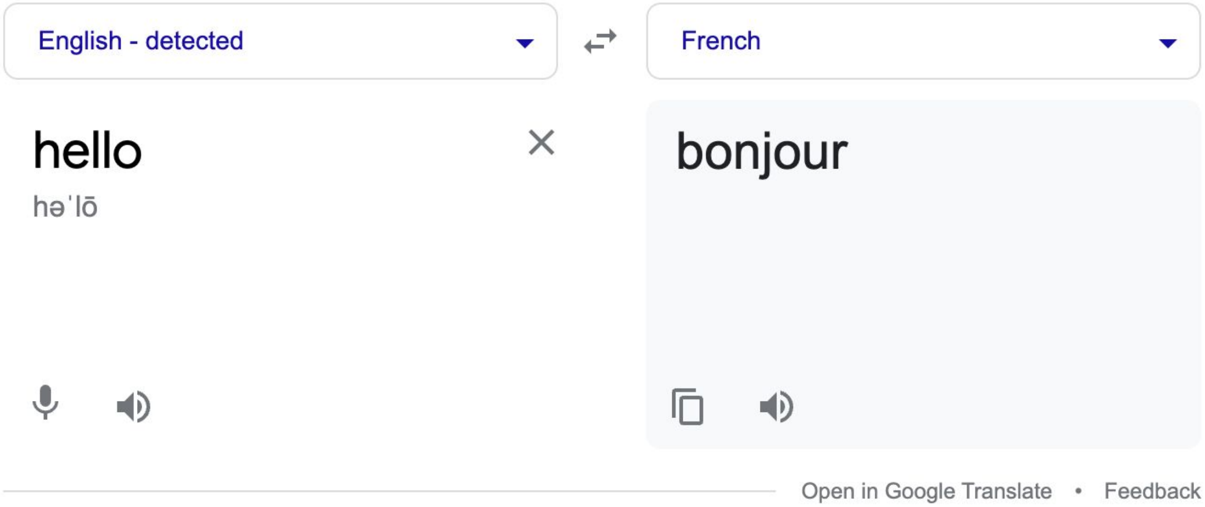
Hannah Brown

Fatemehsadat Mireshghallah

Reza Shokri

Language Models

model language



English - detected

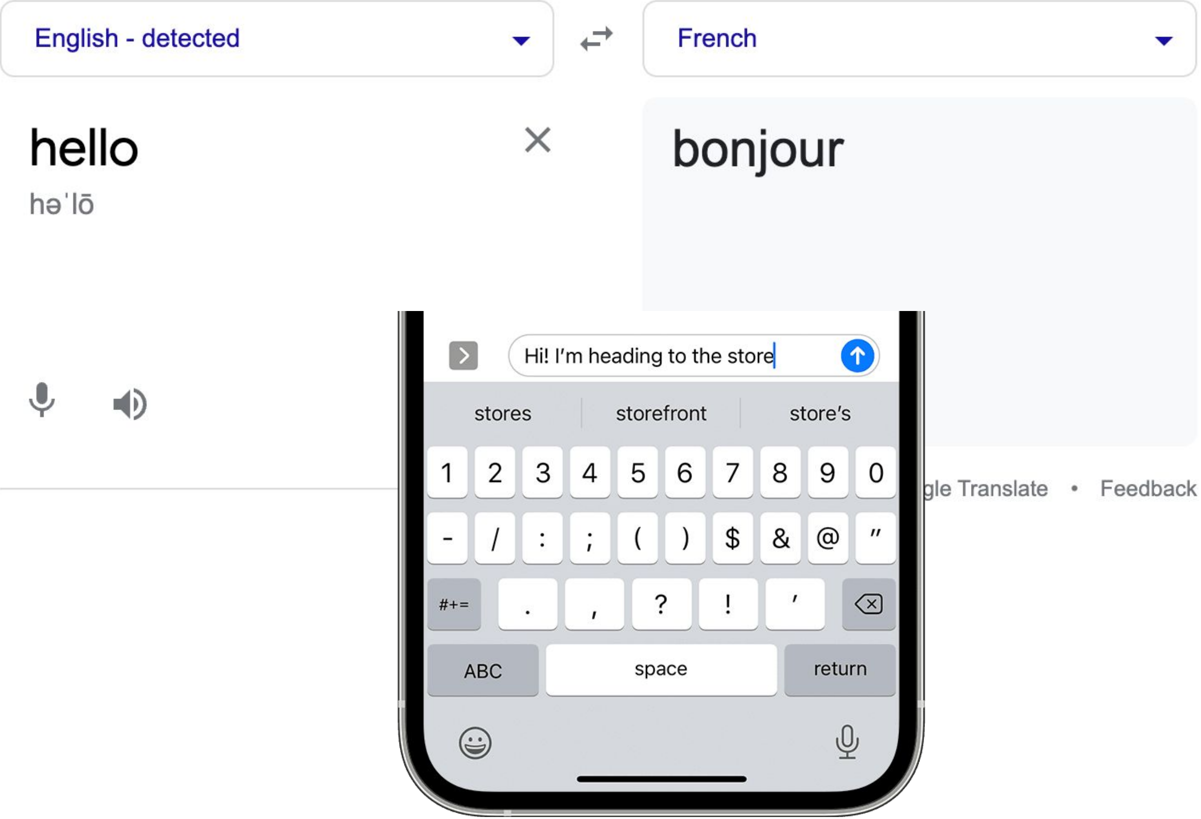
French

hello

həˈlō

bonjour

How to use Auto-Correction and predictive text
Amazon Echo Dot (3rd Gen)



English - detected

French

hello

həˈlō

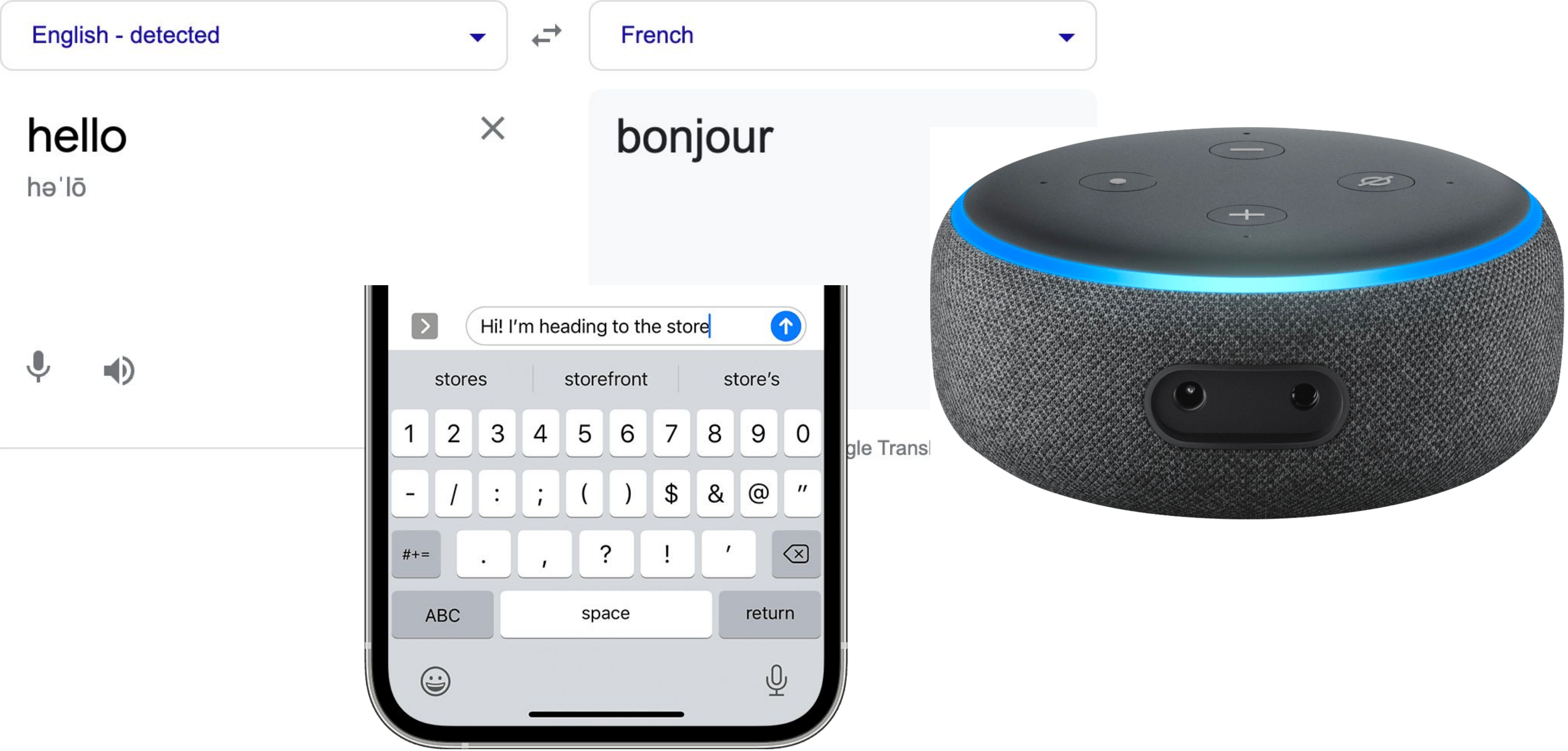
bonjour

gle Translate • Feedback

Hi! I'm heading to the store

stores | storefront | store's

How to use Auto-Correction and predictive text
Amazon Echo Dot (3rd Gen)



Hi! I'm heading to the store

How to use Auto-Correction and predictive text
Amazon Echo Dot (3rd Gen)

Language Models

model language

Language Models

learn a probability distribution of a sequence given the previous tokens

P(word | all previous words)

Language Models

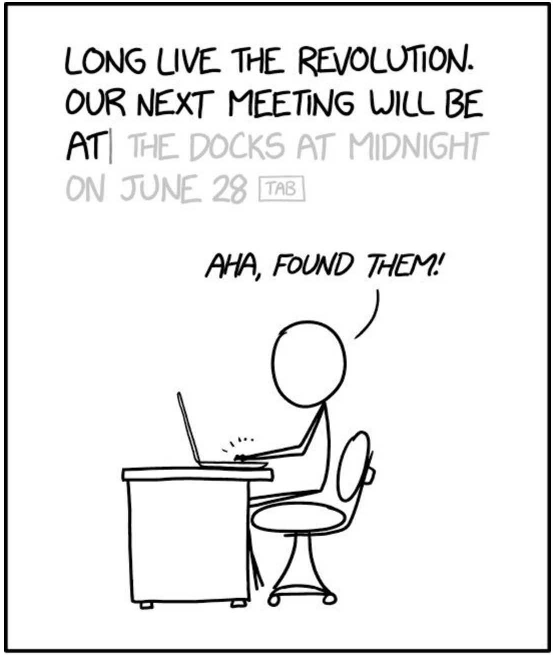
The students opened their ________.

books

laptops

pencils

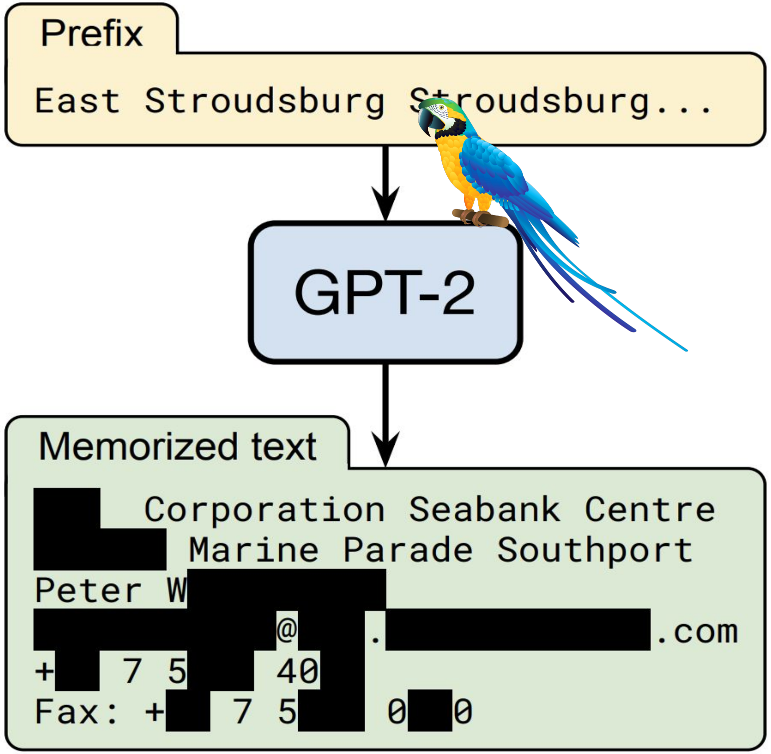
Large Models are Leaky


LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT THE DOCKS AT MIDNIGHT ON JUNE 28
TAB
AHA, FOUND THEM!

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

xkcd.com/2169/

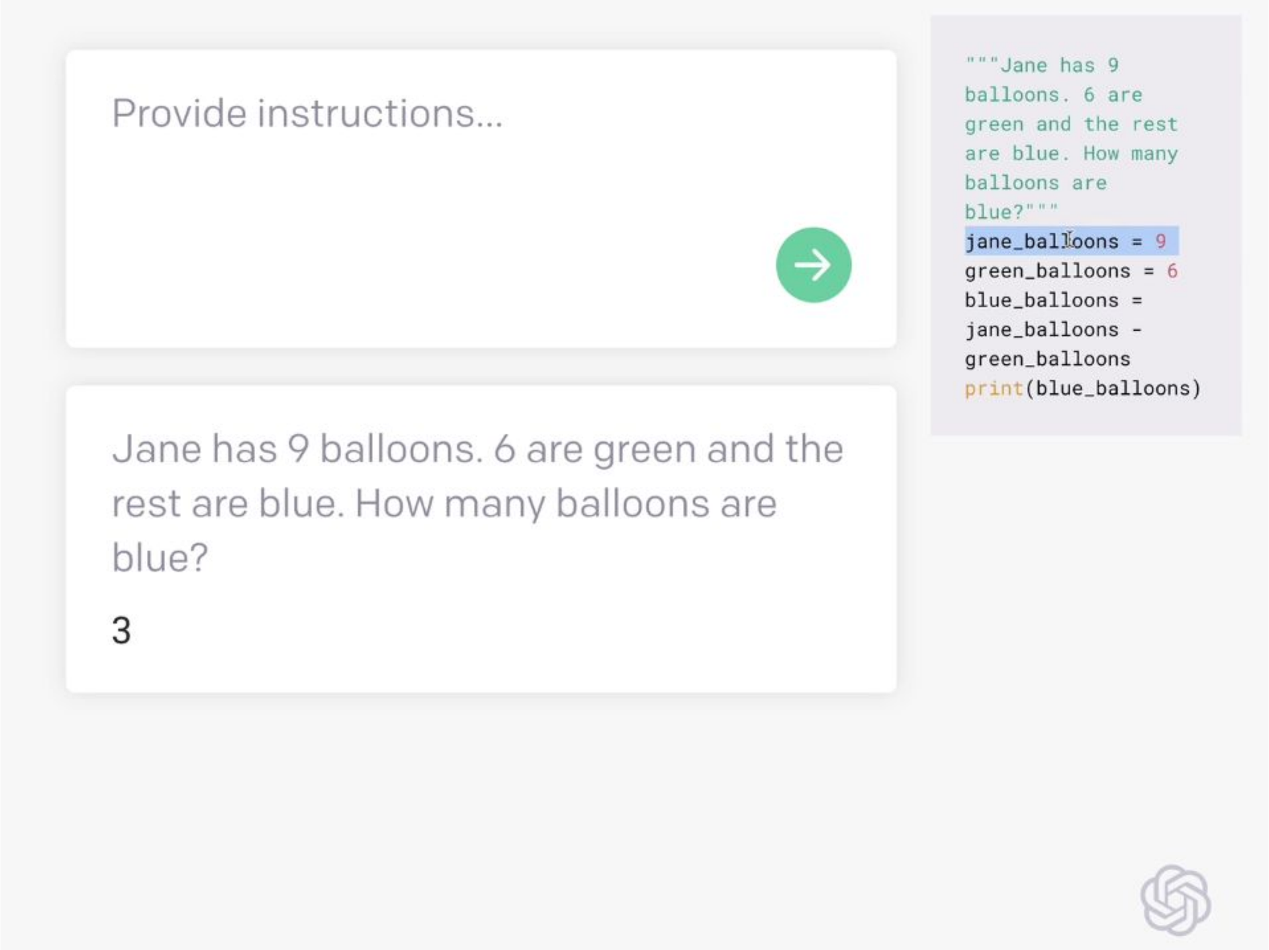
Large Models are Leaky



East Stroudsburg Stroudsburg...

[redacted] Corporation Seabank Centre
Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40
Fax: +[redacted] 7 5[redacted] 0[redacted]0

Carlini et al. Extracting Training Data from Large Language Models. USENIX SEC 2021.

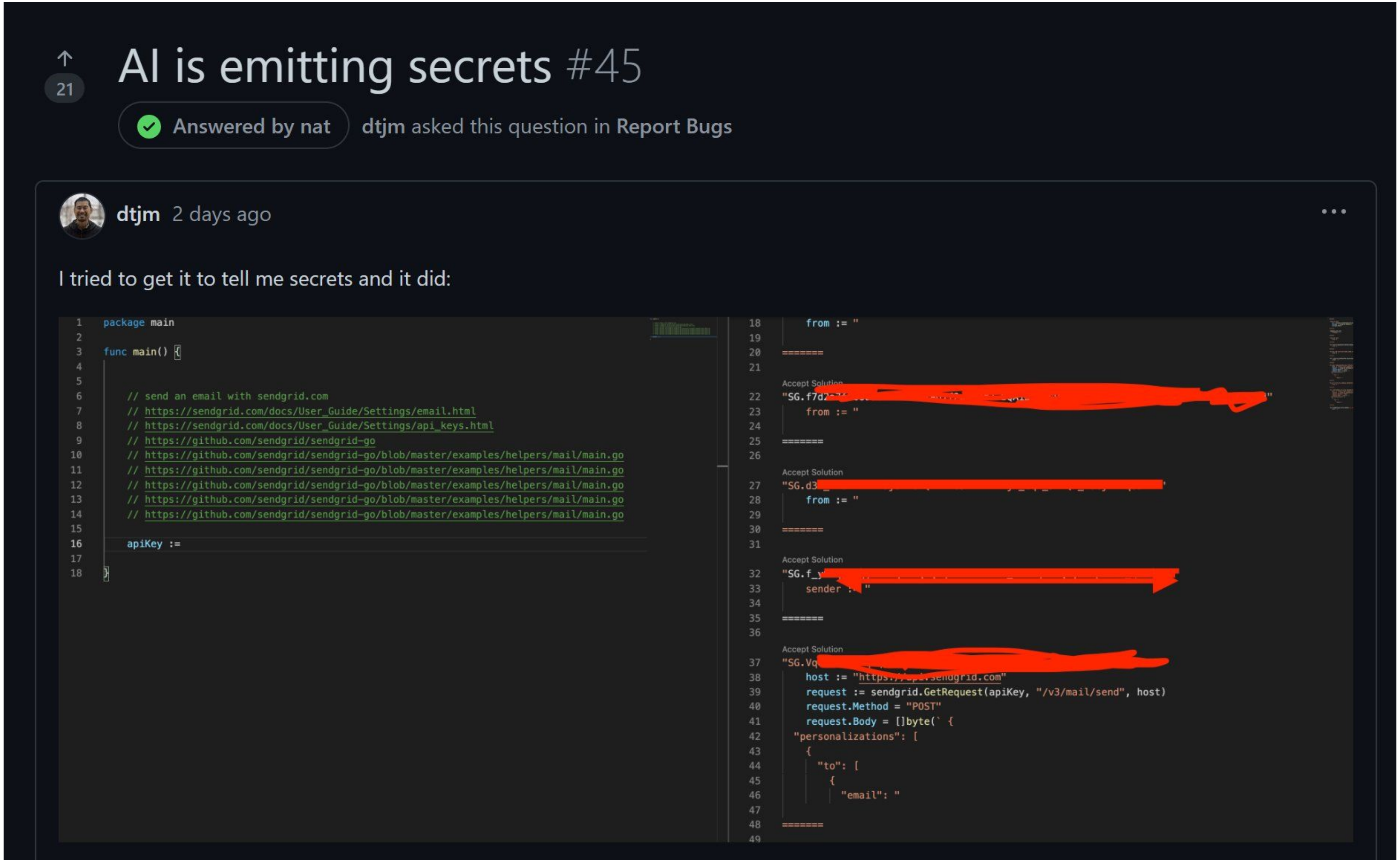


Jane has 9 balloons. 6 are green and the rest are blue. How many balloons are blue?

3

```
"""Jane has 9
balloons. 6 are
green and the rest
are blue. How many
balloons are
blue?"""
jane_balloons = 9
green_balloons = 6
blue_balloons =
jane_balloons -
green_balloons
print(blue_balloons)
```

GitHub Copilot AI Is Generating And Giving Out Functional API Keys



AI is emitting secrets #45

21

Answered by nat · dtjm asked this question in Report Bugs

dtjm 2 days ago

I tried to get it to tell me secrets and it did:

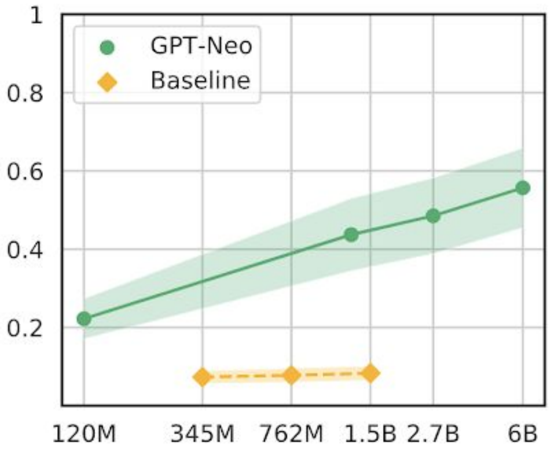
```go
package main

func main() {


    // send an email with sendgrid.com
    // https://sendgrid.com/docs/User_Guide/Settings/email.html
    // https://sendgrid.com/docs/User_Guide/Settings/api_keys.html
    // https://github.com/sendgrid/sendgrid-go
    // https://github.com/sendgrid/sendgrid-go/blob/master/examples/helpers/mail/main.go
    // https://github.com/sendgrid/sendgrid-go/blob/master/examples/helpers/mail/main.go
    // https://github.com/sendgrid/sendgrid-go/blob/master/examples/helpers/mail/main.go
    // https://github.com/sendgrid/sendgrid-go/blob/master/examples/helpers/mail/main.go
    // https://github.com/sendgrid/sendgrid-go/blob/master/examples/helpers/mail/main.go

    apiKey :=

}
```

```
    from := "

=======

Accept Solution
"SG.f7d2[redacted]"
    from := "

=======

Accept Solution
"SG.d3[redacted]
    from := "

=======

Accept Solution
"SG.f_[redacted]
    sender [redacted] "

=======

Accept Solution
"SG.Vq[redacted]
    host := "https[redacted]endgrid.com"
    request := sendgrid.GetRequest(apiKey, "/v3/mail/send", host)
    request.Method = "POST"
    request.Body = []byte(` {
  "personalizations": [
    {
      "to": [
        {
          "email": "

=======
```

GitHub Copilot AI Is Generating And Giving Out Functional API Keys

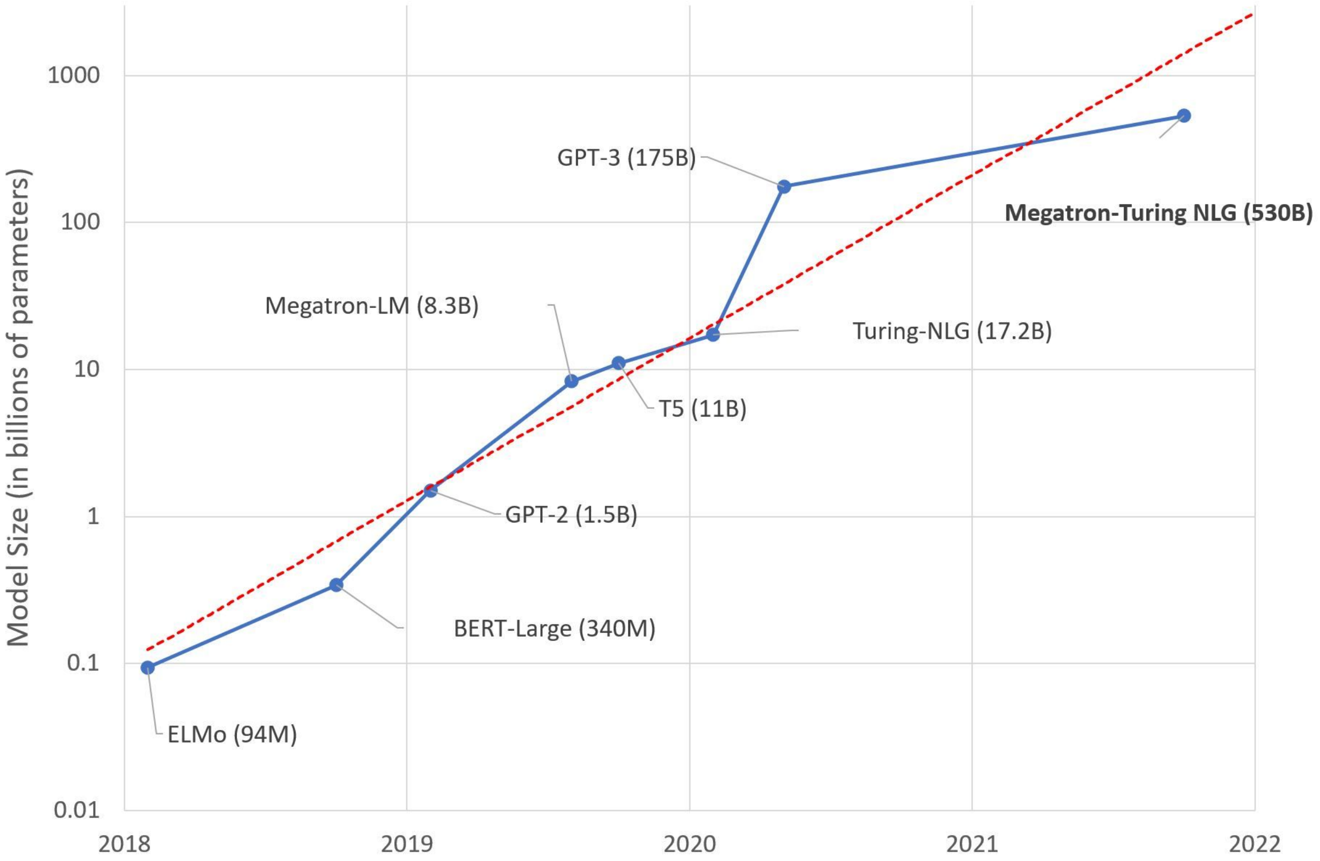
Discoverable memorization **scales**...

...with model scale



(a) Model scale

Carlini, et al., Quantifying Memorization Across Neural Language Models, arxiv, 2022



[Large Language Models: A New Moore's Law?]

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'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

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

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.

Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022,

Neural language models memorize training data.

571x

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.

51x

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

5,497x

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 expert 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

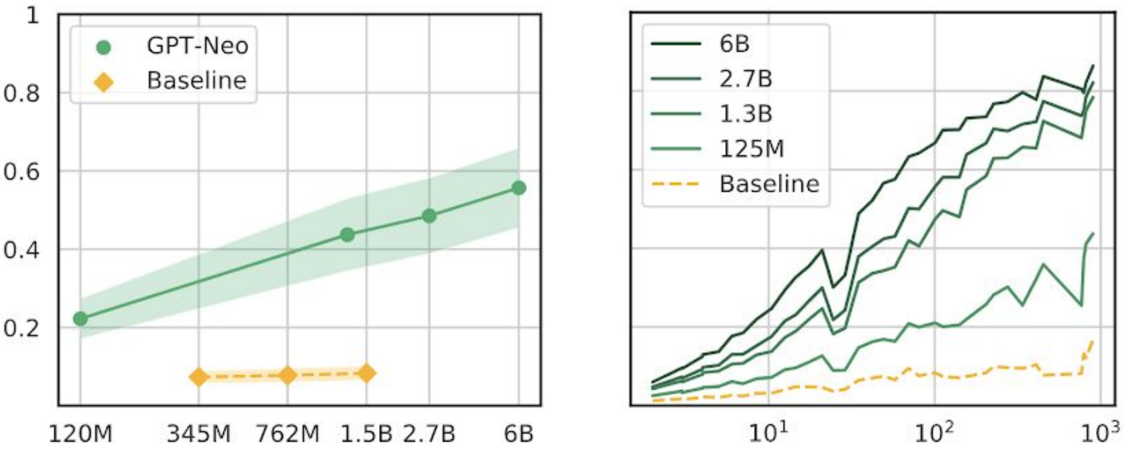
6x

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.

Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022,

Discoverable memorization **scales**...

...with data repetition



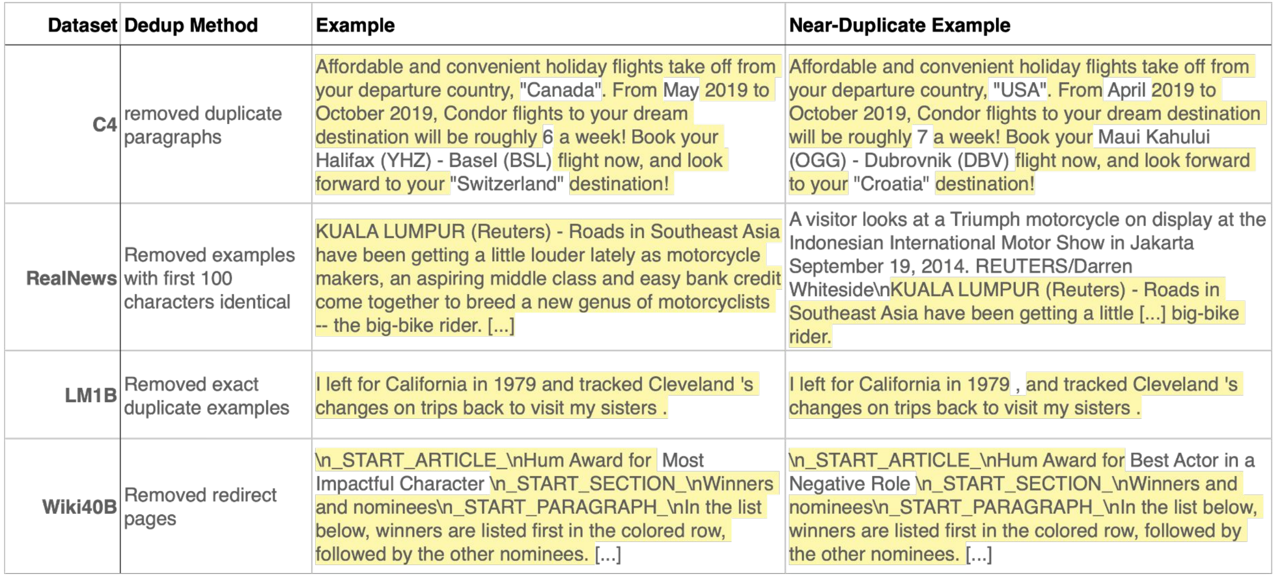
(a) Model scale

(b) Data repetition

Carlini, et al., Quantifying Memorization Across Neural Language Models, arxiv, 2022

Thorough deduplication is hard.

Existing datasets are insufficiently deduplicated.



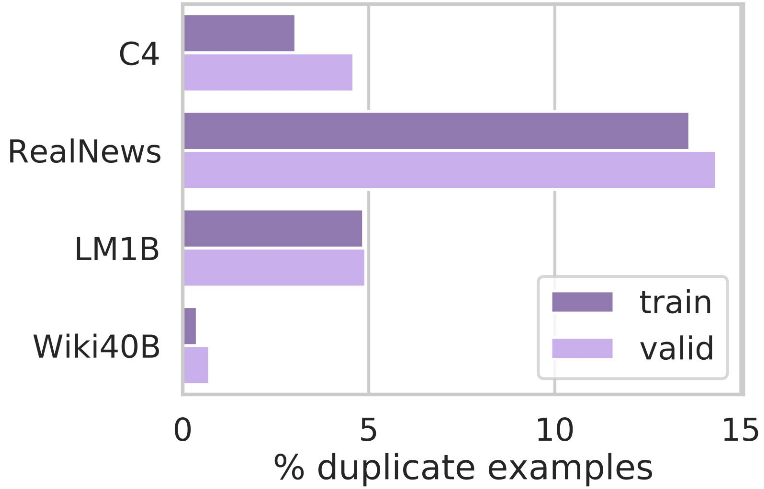
Dataset	Dedup Method	Example	Near-Duplicate Example
C4	removed duplicate paragraphs	Affordable and convenient holiday flights take off from your departure country, "Canada". From May 2019 to October 2019, Condor flights to your dream destination will be roughly 6 a week! Book your Halifax (YHZ) - Basel (BSL) flight now, and look forward to your "Switzerland" destination!	Affordable and convenient holiday flights take off from your departure country, "USA". From April 2019 to October 2019, Condor flights to your dream destination will be roughly 7 a week! Book your Maui Kahului (OGG) - Dubrovnik (DBV) flight now, and look forward to your "Croatia" destination!
RealNews	Removed examples with first 100 characters identical	KUALA LUMPUR (Reuters) - Roads in Southeast Asia have been getting a little louder lately as motorcycle makers, an aspiring middle class and easy bank credit come together to breed a new genus of motorcyclists -- the big-bike rider. [...]	A visitor looks at a Triumph motorcycle on display at the Indonesian International Motor Show in Jakarta September 19, 2014. REUTERS/Darren Whiteside\nKUALA LUMPUR (Reuters) - Roads in Southeast Asia have been getting a little [...] big-bike rider.
LM1B	Removed exact duplicate examples	I left for California in 1979 and tracked Cleveland 's changes on trips back to visit my sisters .	I left for California in 1979 , and tracked Cleveland 's changes on trips back to visit my sisters .
Wiki40B	Removed redirect pages	\n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]	\n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]

Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022.

Deduplicating Text Data

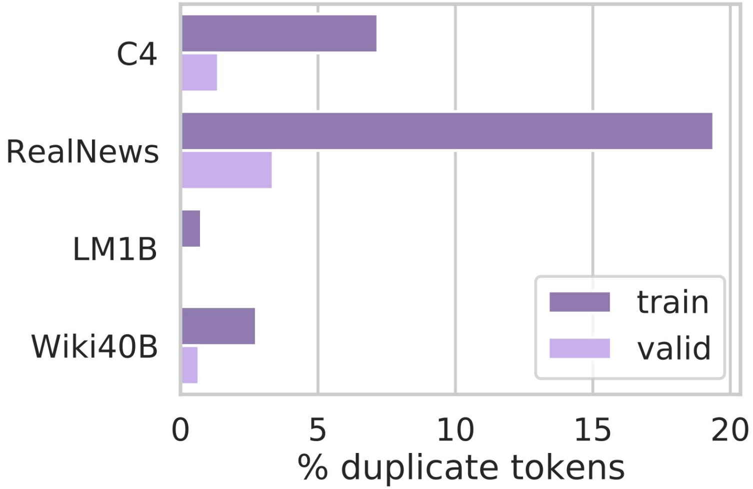
Near Duplicates

Cluster examples with high n-gram overlap using MinHash.
Delete all but one example from each cluster.



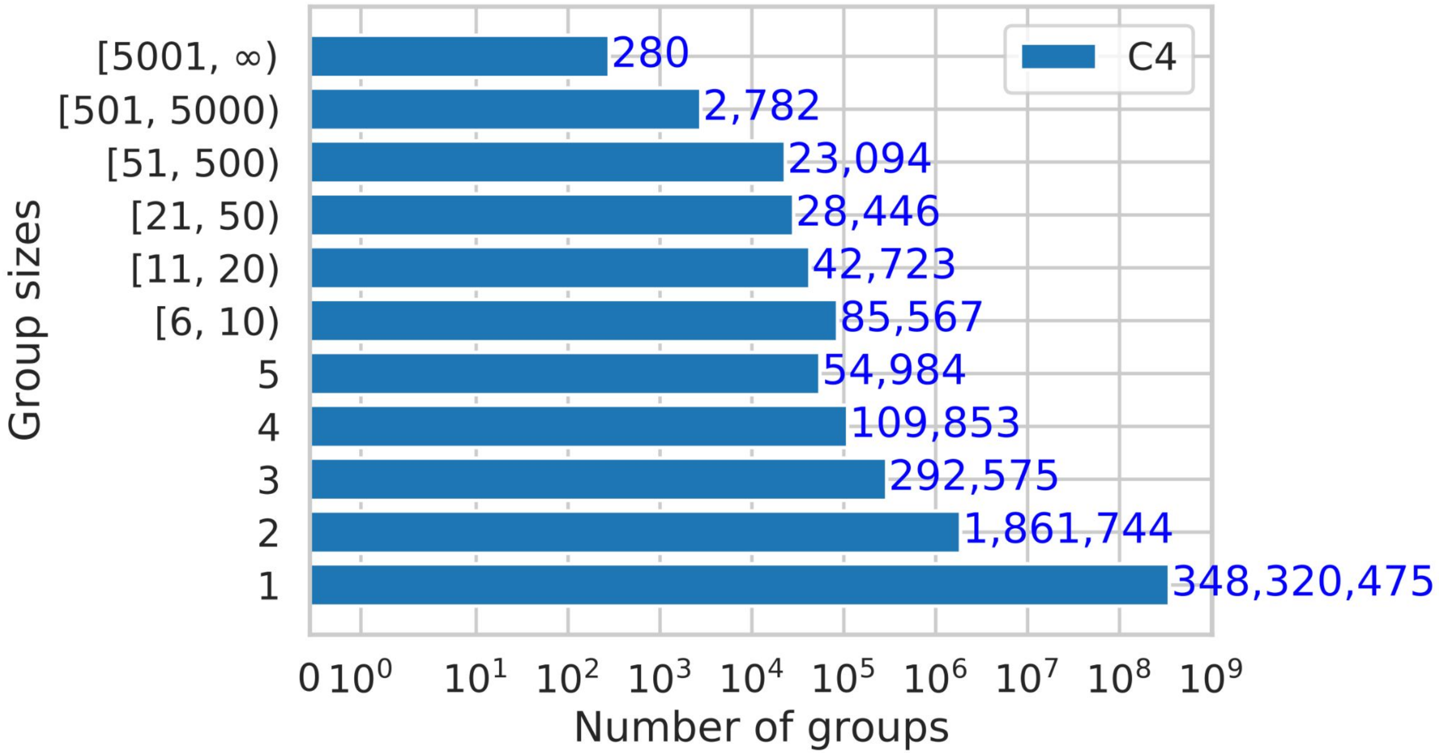
Exact Substring

Insert dataset into suffix array.
Delete repeated 50-token substrings.



Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022,

C4 Near-Duplicate Clusters



Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022.

Experimental Design

Train

Three 1.5B decoder-only LMs on:

Original C4

C4 deduplicated with NearDup

C4 deduplicated with ExactSubstr

Measure

Prompted memorization

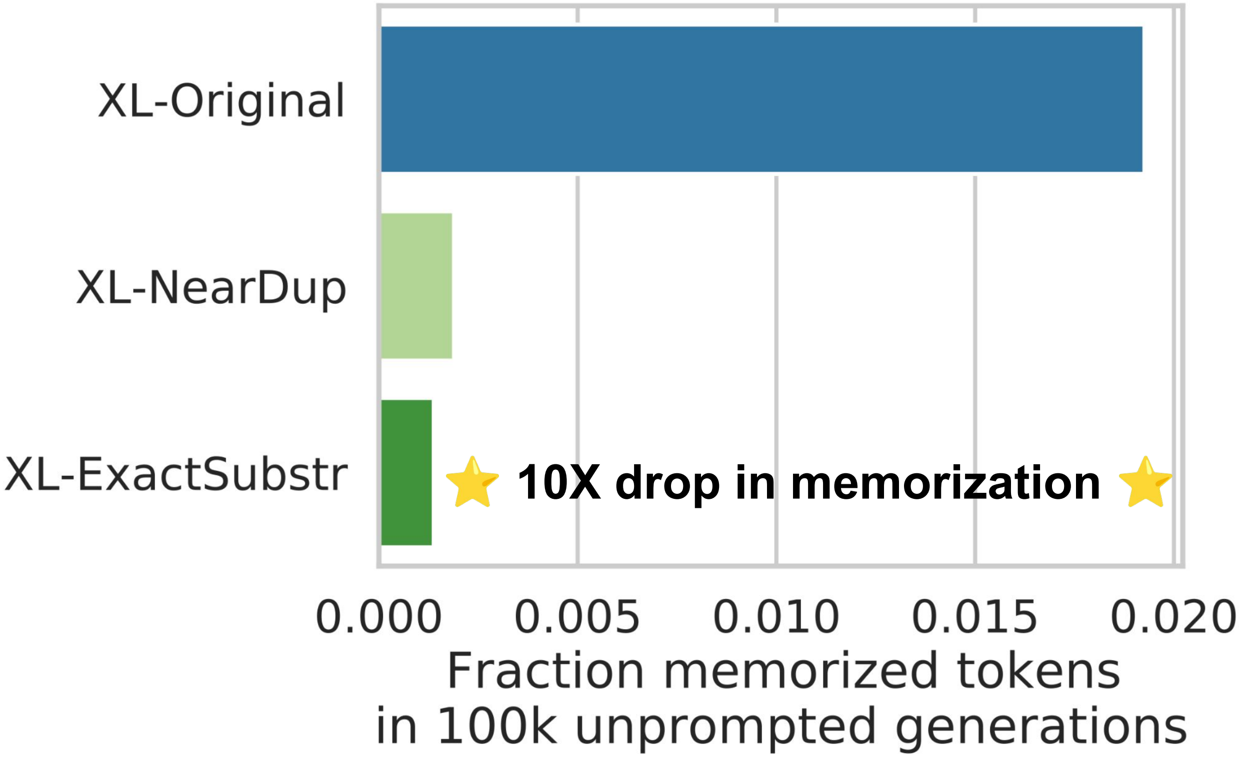
Unprompted memorization

Perplexity on evaluation datasets.

Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022,

Unprompted Memorization

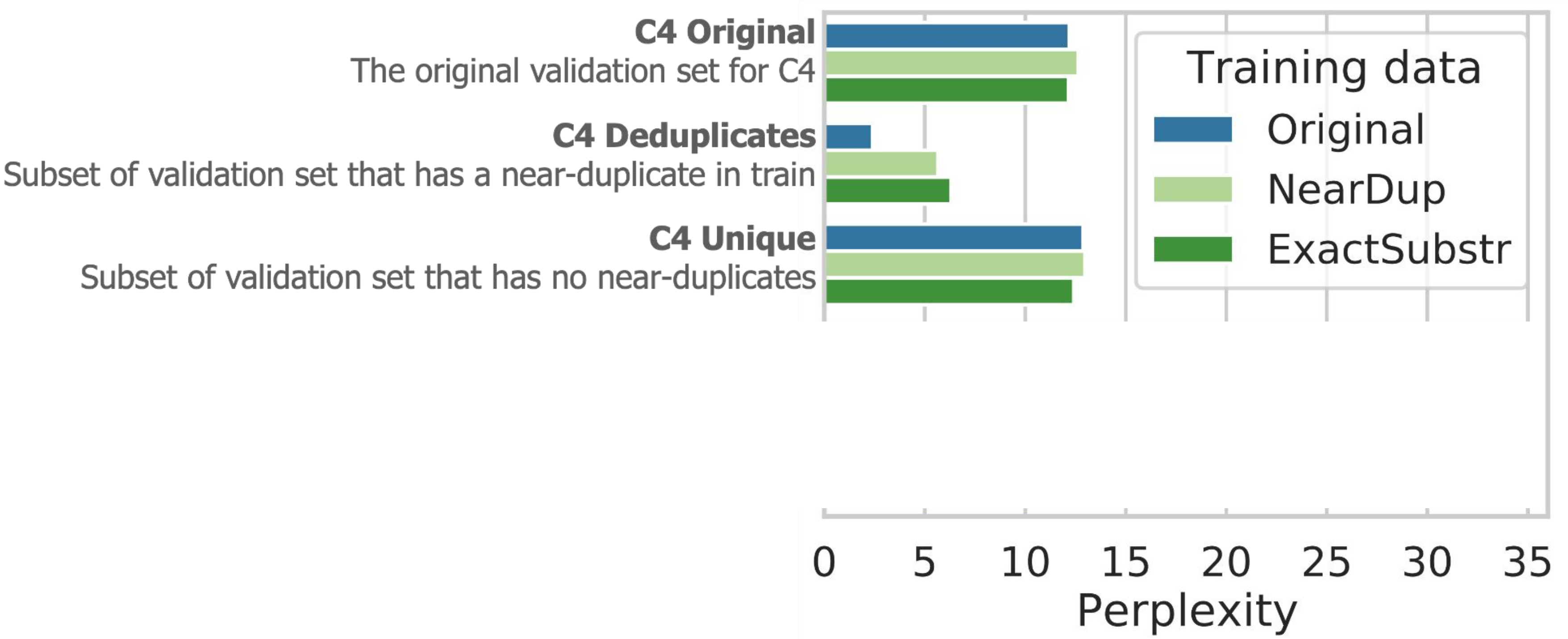
Generate text with random sampling (top-k=50)



⭐ **10X drop in memorization** ⭐

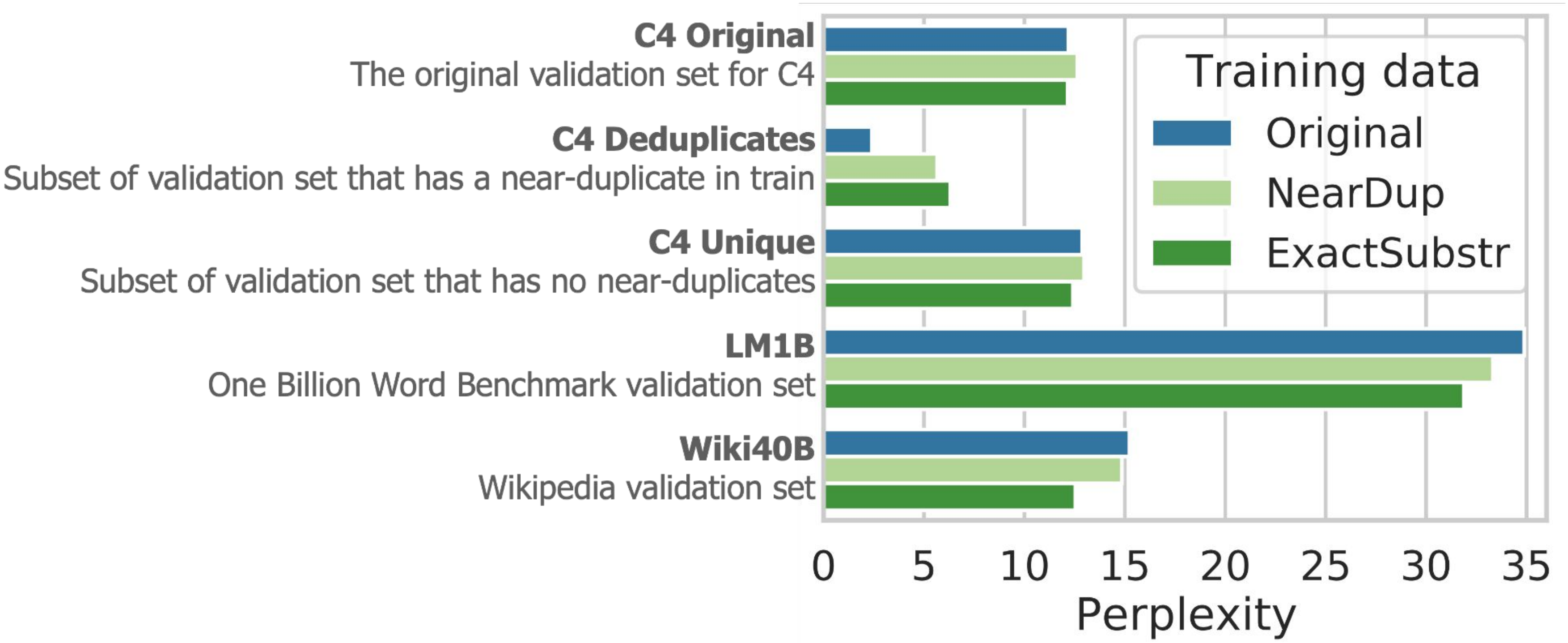
Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022.

Deduplicated models are better.



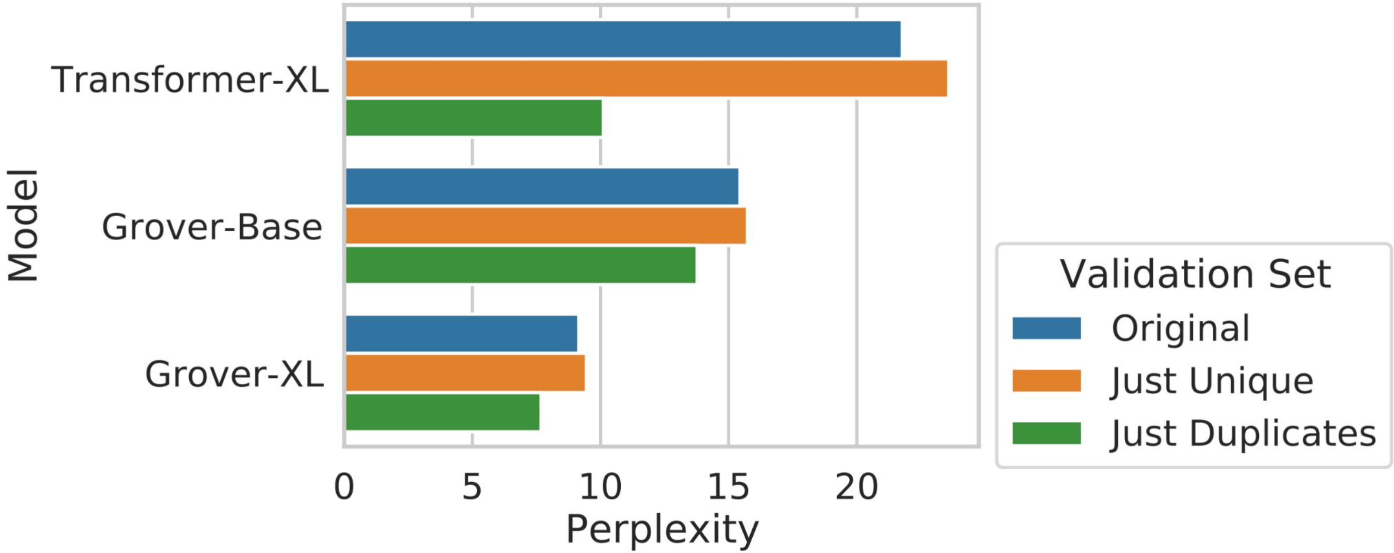
Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022.

Deduplicated models are better.



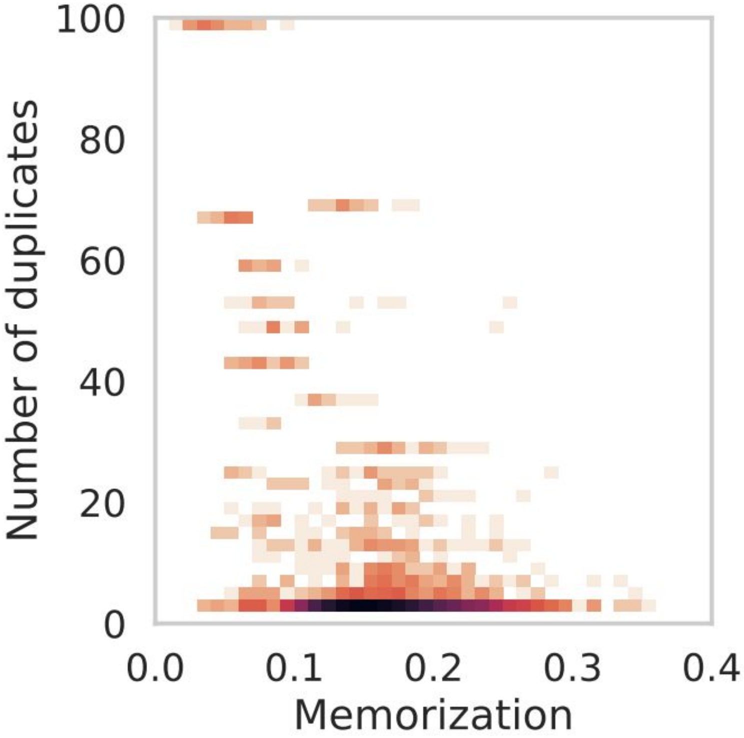
Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022.

Train-test leakage harms evaluation.



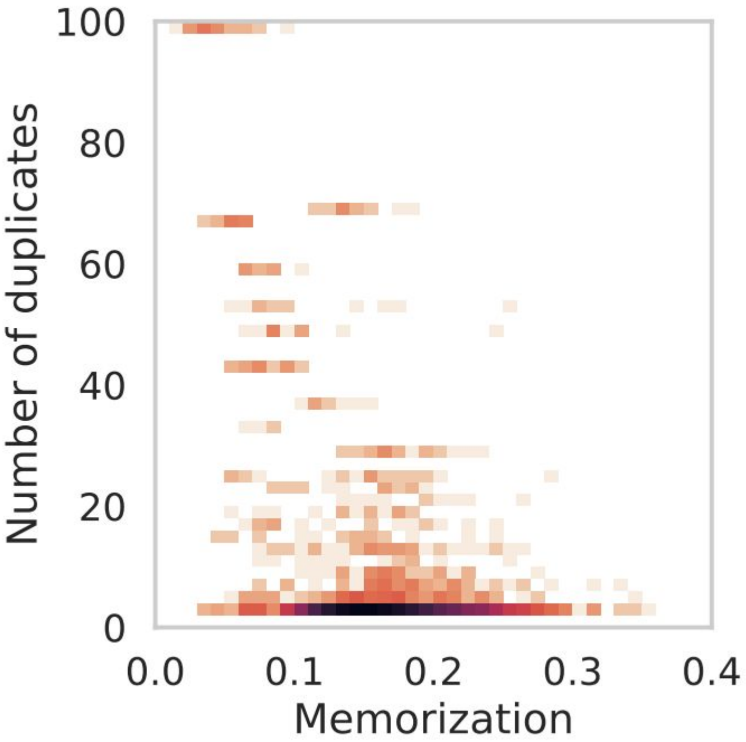
Can we find examples that are rarely seen but still memorized?

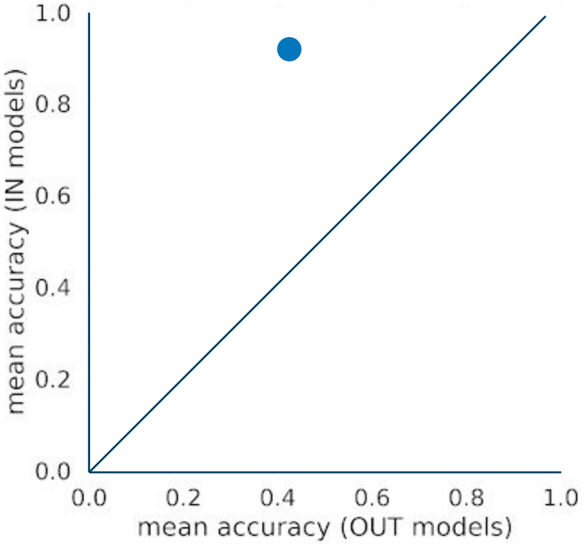
Counterfactual Memorization



Zhang, et al., Counterfactual Memorization in Neural Language Models, arxiv, 2022

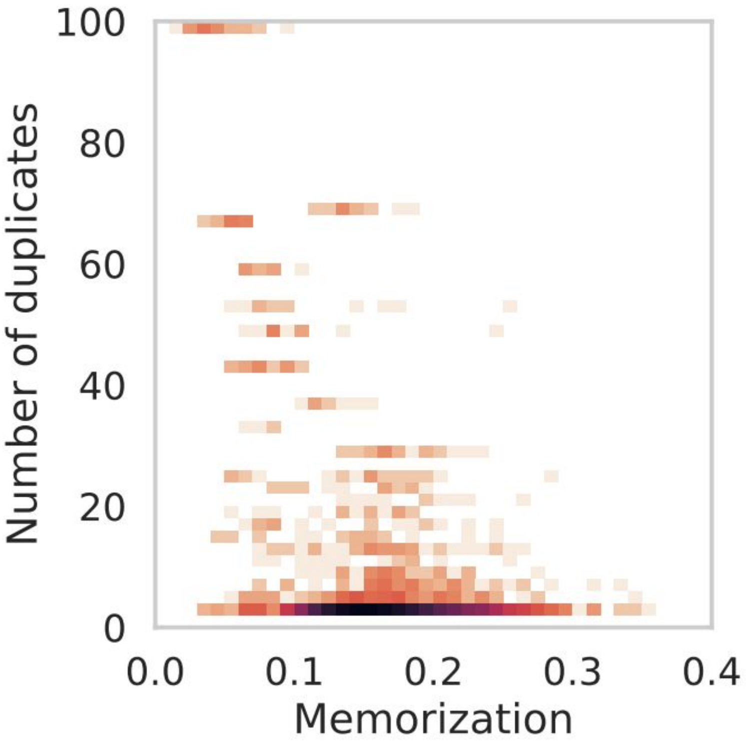
Counterfactual Memorization

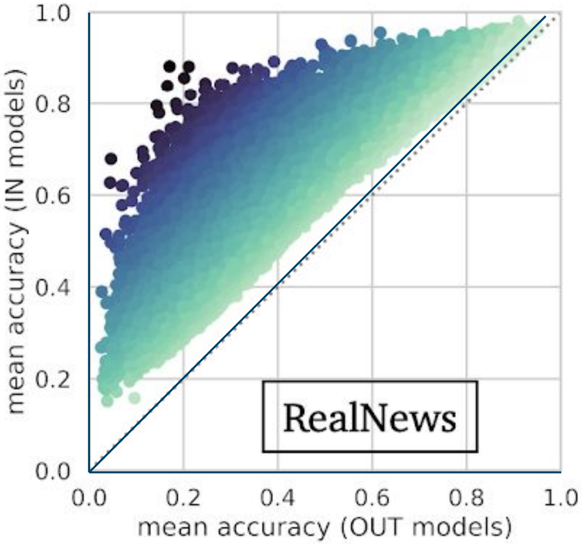




Zhang, et al., Counterfactual Memorization in Neural Language Models, arxiv, 2022

Counterfactual Memorization

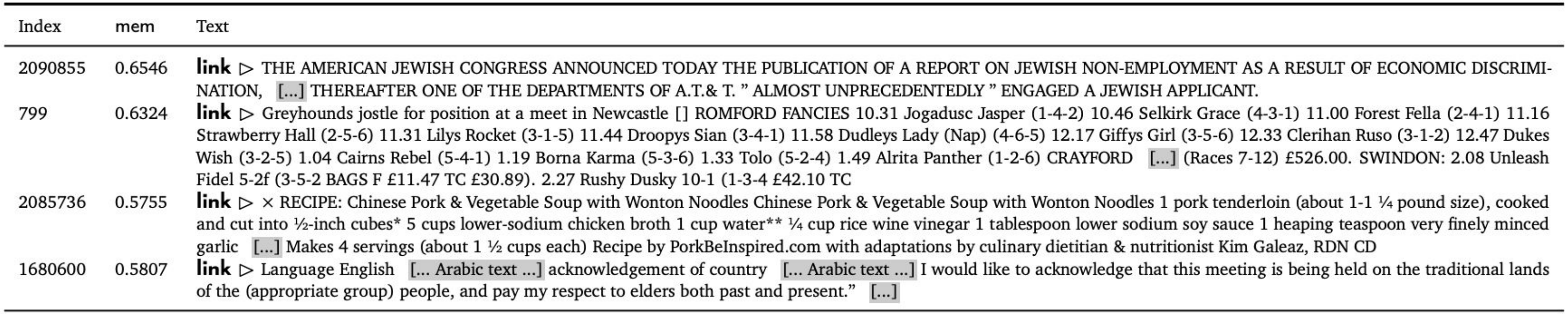




Zhang, et al., Counterfactual Memorization in Neural Language Models, arxiv, 2022

Some examples

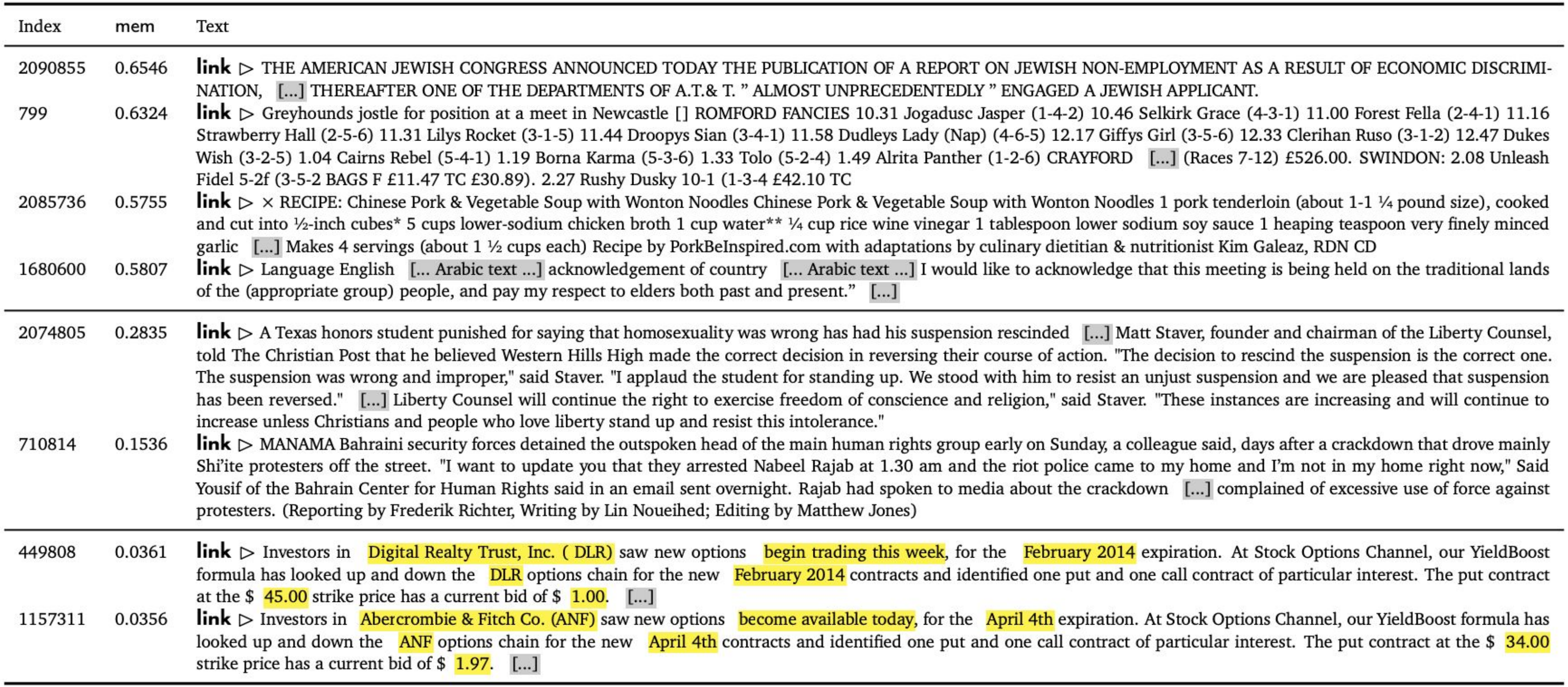
high



Index	mem	Text
2090855	0.6546	link ▷ THE AMERICAN JEWISH CONGRESS ANNOUNCED TODAY THE PUBLICATION OF A REPORT ON JEWISH NON-EMPLOYMENT AS A RESULT OF ECONOMIC DISCRIMINATION, [...] THEREAFTER ONE OF THE DEPARTMENTS OF A.T.& T. " ALMOST UNPRECEDENTEDLY " ENGAGED A JEWISH APPLICANT.
799	0.6324	link ▷ Greyhounds jostle for position at a meet in Newcastle [] ROMFORD FANCIES 10.31 Jogadusc Jasper (1-4-2) 10.46 Selkirk Grace (4-3-1) 11.00 Forest Fella (2-4-1) 11.16 Strawberry Hall (2-5-6) 11.31 Lilys Rocket (3-1-5) 11.44 Droopys Sian (3-4-1) 11.58 Dudleys Lady (Nap) (4-6-5) 12.17 Giffys Girl (3-5-6) 12.33 Clerihan Ruso (3-1-2) 12.47 Dukes Wish (3-2-5) 1.04 Cairns Rebel (5-4-1) 1.19 Borna Karma (5-3-6) 1.33 Tolo (5-2-4) 1.49 Alrita Panther (1-2-6) CRAYFORD [...] (Races 7-12) £526.00. SWINDON: 2.08 Unleash Fidel 5-2f (3-5-2 BAGS F £11.47 TC £30.89). 2.27 Rushy Dusky 10-1 (1-3-4 £42.10 TC
2085736	0.5755	link ▷ × RECIPE: Chinese Pork & Vegetable Soup with Wonton Noodles Chinese Pork & Vegetable Soup with Wonton Noodles 1 pork tenderloin (about 1-1 ¼ pound size), cooked and cut into ½-inch cubes* 5 cups lower-sodium chicken broth 1 cup water** ¼ cup rice wine vinegar 1 tablespoon lower sodium soy sauce 1 heaping teaspoon very finely minced garlic [...] Makes 4 servings (about 1 ½ cups each) Recipe by PorkBeInspired.com with adaptations by culinary dietitian & nutritionist Kim Galeaz, RDN CD
1680600	0.5807	link ▷ Language English [... Arabic text ...] acknowledgement of country [... Arabic text ...] I would like to acknowledge that this meeting is being held on the traditional lands of the (appropriate group) people, and pay my respect to elders both past and present." [...]

Zhang, et al., Counterfactual Memorization in Neural Language Models, arxiv, 2022

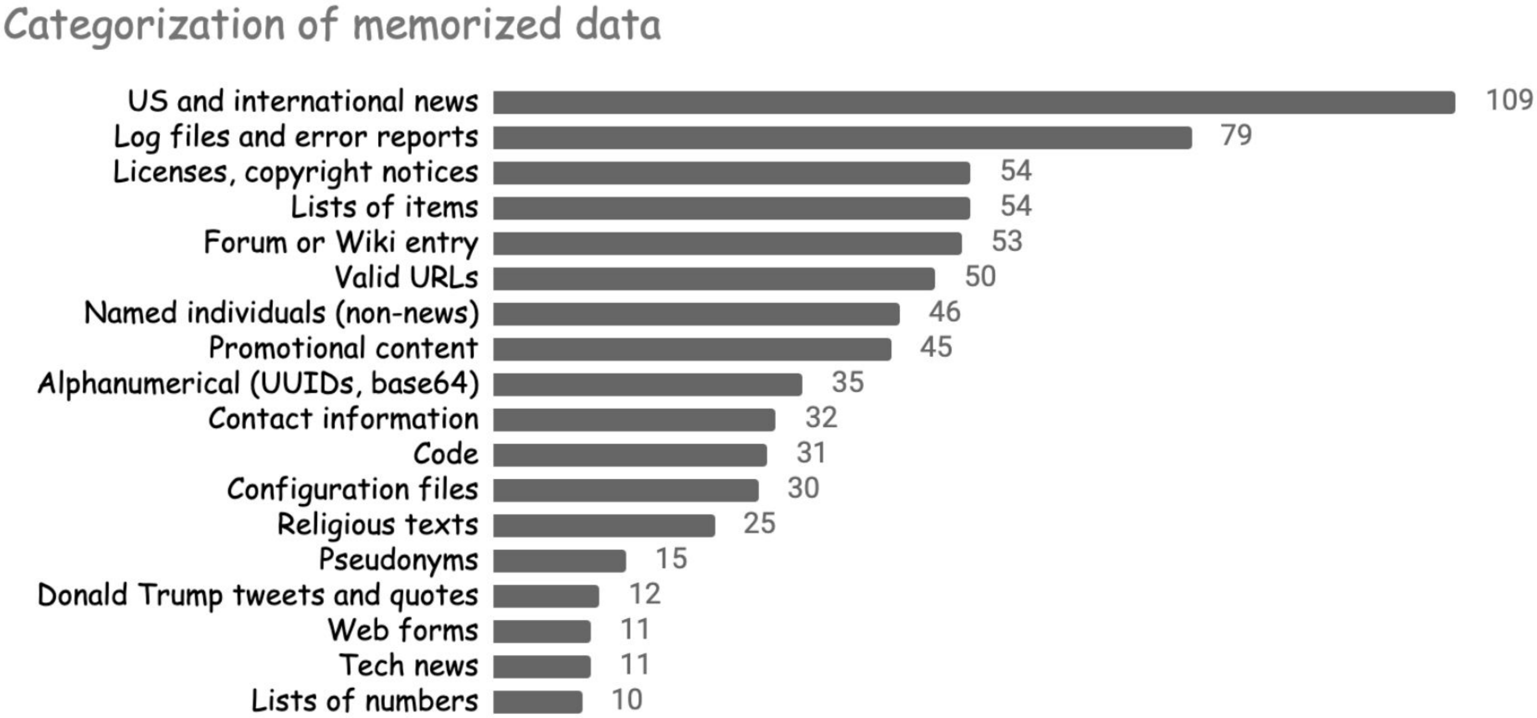
Some examples



	Index	mem	Text
high	2090855	0.6546	**link** ▷ THE AMERICAN JEWISH CONGRESS ANNOUNCED TODAY THE PUBLICATION OF A REPORT ON JEWISH NON-EMPLOYMENT AS A RESULT OF ECONOMIC DISCRIMINATION, [...] THEREAFTER ONE OF THE DEPARTMENTS OF A.T.& T. " ALMOST UNPRECEDENTEDLY " ENGAGED A JEWISH APPLICANT.
	799	0.6324	**link** ▷ Greyhounds jostle for position at a meet in Newcastle [] ROMFORD FANCIES 10.31 Jogadusc Jasper (1-4-2) 10.46 Selkirk Grace (4-3-1) 11.00 Forest Fella (2-4-1) 11.16 Strawberry Hall (2-5-6) 11.31 Lilys Rocket (3-1-5) 11.44 Droopys Sian (3-4-1) 11.58 Dudleys Lady (Nap) (4-6-5) 12.17 Giffys Girl (3-5-6) 12.33 Clerihan Ruso (3-1-2) 12.47 Dukes Wish (3-2-5) 1.04 Cairns Rebel (5-4-1) 1.19 Borna Karma (5-3-6) 1.33 Tolo (5-2-4) 1.49 Alrita Panther (1-2-6) CRAYFORD [...] (Races 7-12) £526.00. SWINDON: 2.08 Unleash Fidel 5-2f (3-5-2 BAGS F £11.47 TC £30.89). 2.27 Rushy Dusky 10-1 (1-3-4 £42.10 TC
	2085736	0.5755	**link** ▷ × RECIPE: Chinese Pork & Vegetable Soup with Wonton Noodles Chinese Pork & Vegetable Soup with Wonton Noodles 1 pork tenderloin (about 1-1 ¼ pound size), cooked and cut into ½-inch cubes* 5 cups lower-sodium chicken broth 1 cup water** ¼ cup rice wine vinegar 1 tablespoon lower sodium soy sauce 1 heaping teaspoon very finely minced garlic [...] Makes 4 servings (about 1 ½ cups each) Recipe by PorkBeInspired.com with adaptations by culinary dietitian & nutritionist Kim Galeaz, RDN CD
	1680600	0.5807	**link** ▷ Language English [... Arabic text ...] acknowledgement of country [... Arabic text ...] I would like to acknowledge that this meeting is being held on the traditional lands of the (appropriate group) people, and pay my respect to elders both past and present." [...]
med	2074805	0.2835	**link** ▷ A Texas honors student punished for saying that homosexuality was wrong has had his suspension rescinded [...] Matt Staver, founder and chairman of the Liberty Counsel, told The Christian Post that he believed Western Hills High made the correct decision in reversing their course of action. "The decision to rescind the suspension is the correct one. The suspension was wrong and improper," said Staver. "I applaud the student for standing up. We stood with him to resist an unjust suspension and we are pleased that suspension has been reversed." [...] Liberty Counsel will continue the right to exercise freedom of conscience and religion," said Staver. "These instances are increasing and will continue to increase unless Christians and people who love liberty stand up and resist this intolerance."
	710814	0.1536	**link** ▷ MANAMA Bahraini security forces detained the outspoken head of the main human rights group early on Sunday, a colleague said, days after a crackdown that drove mainly Shi'ite protesters off the street. "I want to update you that they arrested Nabeel Rajab at 1.30 am and the riot police came to my home and I'm not in my home right now," Said Yousif of the Bahrain Center for Human Rights said in an email sent overnight. Rajab had spoken to media about the crackdown [...] complained of excessive use of force against protesters. (Reporting by Frederik Richter, Writing by Lin Noueihed; Editing by Matthew Jones)
low	449808	0.0361	**link** ▷ Investors in Digital Realty Trust, Inc. (DLR) saw new options begin trading this week, for the February 2014 expiration. At Stock Options Channel, our YieldBoost formula has looked up and down the DLR options chain for the new February 2014 contracts and identified one put and one call contract of particular interest. The put contract at the $ 45.00 strike price has a current bid of $ 1.00. [...]
	1157311	0.0356	**link** ▷ Investors in Abercrombie & Fitch Co. (ANF) saw new options become available today, for the April 4th expiration. At Stock Options Channel, our YieldBoost formula has looked up and down the ANF options chain for the new April 4th contracts and identified one put and one call contract of particular interest. The put contract at the $ 34.00 strike price has a current bid of $ 1.97. [...]

Zhang, et al., Counterfactual Memorization in Neural Language Models, arxiv, 2022

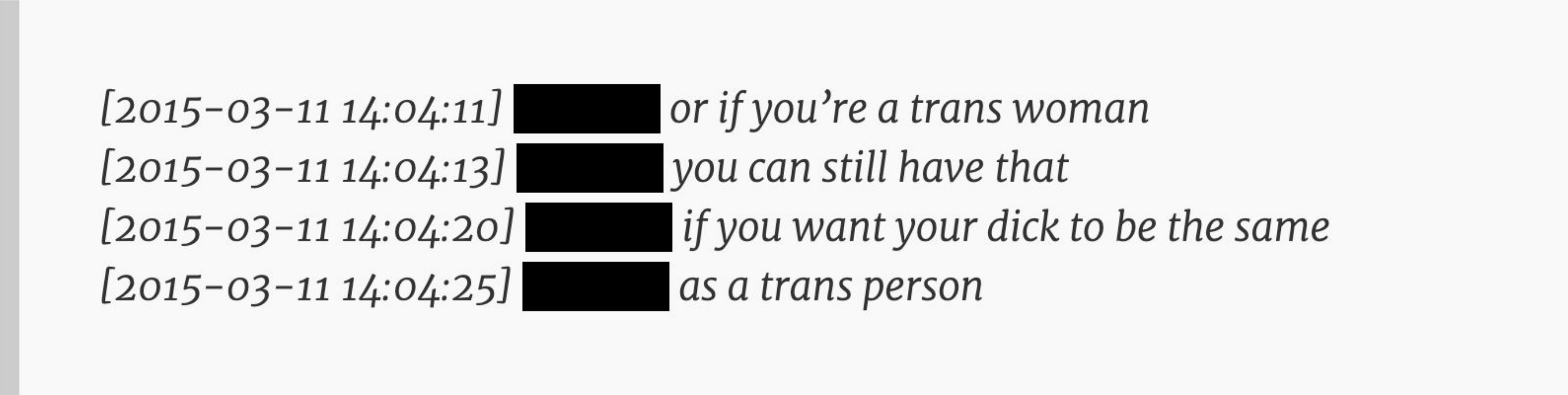
What sort of examples get memorized?



Carlini, et al., Quantifying Memorization Across Neural Language Models, arxiv, 2022

Chat logs

Fictitious generated example based on a real conversation

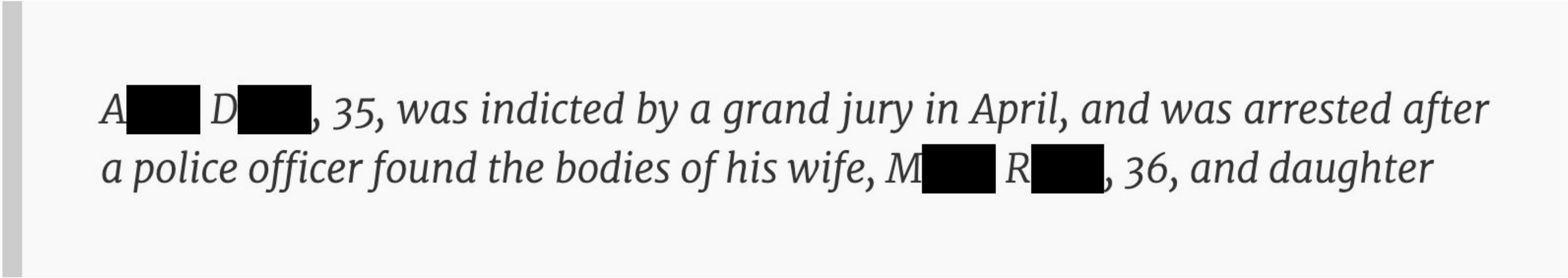


```
[2015-03-11 14:04:11] [redacted] or if you're a trans woman
[2015-03-11 14:04:13] [redacted] you can still have that
[2015-03-11 14:04:20] [redacted] if you want your dick to be the same
[2015-03-11 14:04:25] [redacted] as a trans person
```

Does GPT-2 Know Your Phone Number?

News Articles

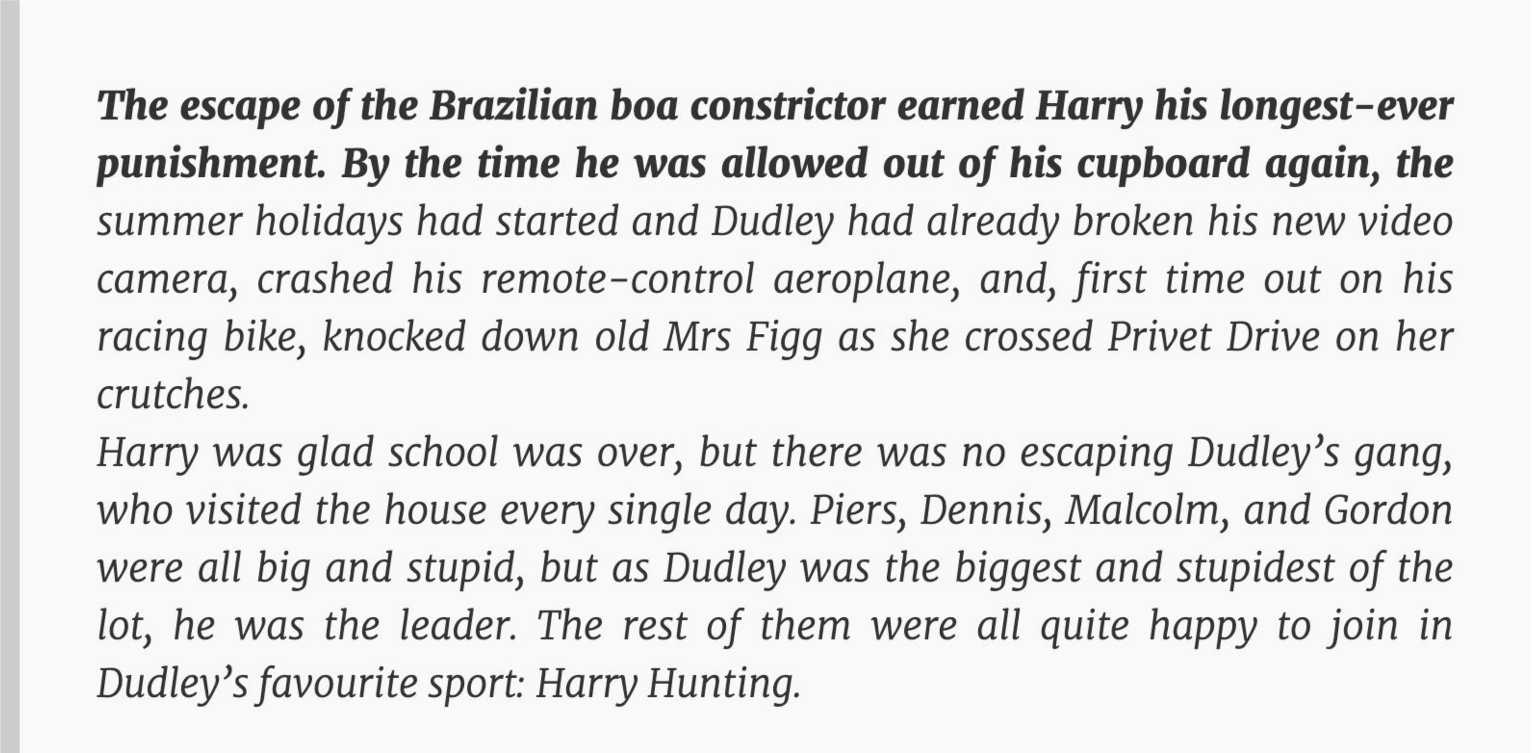
Misidentifying individuals in news articles



A[redacted] D[redacted], 35, was indicted by a grand jury in April, and was arrested after a police officer found the bodies of his wife, M[redacted] R[redacted], 36, and daughter

Does GPT-2 Know Your Phone Number?

Copyrighted text

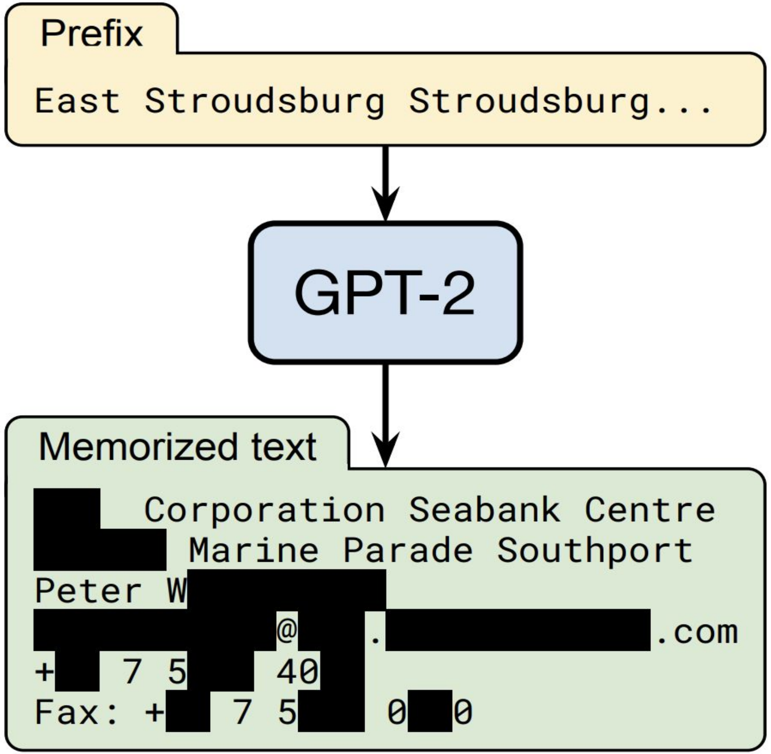


The escape of the Brazilian boa constrictor earned Harry his longest-ever punishment. By the time he was allowed out of his cupboard again, the *summer holidays had started and Dudley had already broken his new video camera, crashed his remote-control aeroplane, and, first time out on his racing bike, knocked down old Mrs Figg as she crossed Privet Drive on her crutches.*

Harry was glad school was over, but there was no escaping Dudley's gang, who visited the house every single day. Piers, Dennis, Malcolm, and Gordon were all big and stupid, but as Dudley was the biggest and stupidest of the lot, he was the leader. The rest of them were all quite happy to join in Dudley's favourite sport: Harry Hunting.

Does GPT-2 Know Your Phone Number?

Large Models are Leaky



Prefix

East Stroudsburg Stroudsburg...

GPT-2

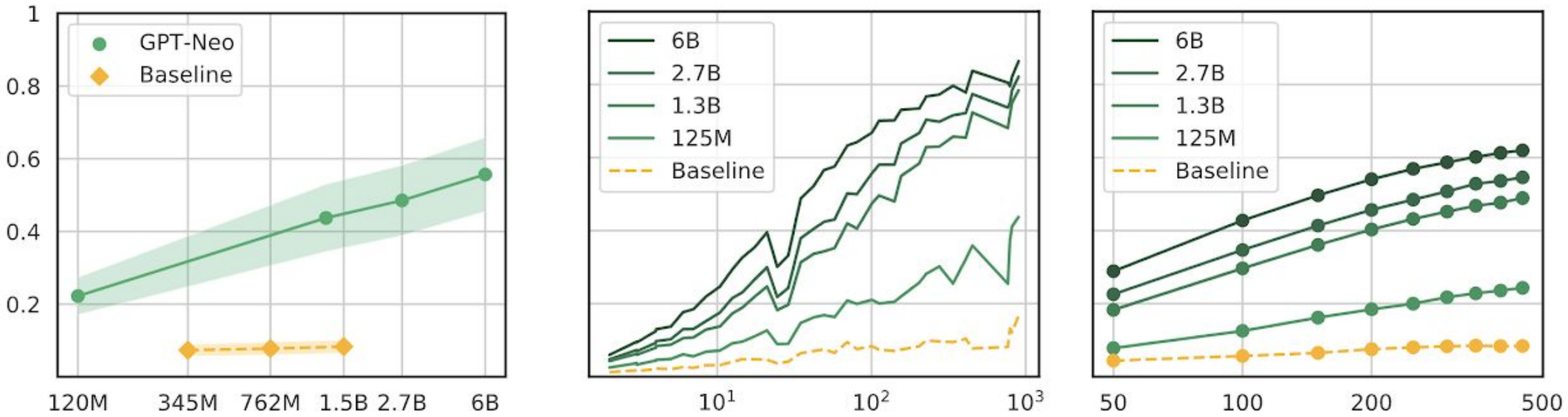
Memorized text

[redacted] Corporation Seabank Centre
Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0

Carlini, et al., Quantifying Memorization Across Neural Language Models, arxiv, 2022

Discoverable memorization **scales**…

…with context length



(a) Model scale

(b) Data repetition

(c) Context size

Carlini, et al., Quantifying Memorization Across Neural Language Models, arxiv, 2022

Harm from memorization is contextual

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

Facts: “Christmas is celebrated on Dec 25th”

Harm from memorization is contextual

Common phrases: “To whom it may concern…”

Facts: “Christmas is celebrated on Dec 25th”

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

Harm from memorization is contextual

Common phrases: “To whom it may concern…”

Facts: “Christmas is celebrated on Dec 25th”

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

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

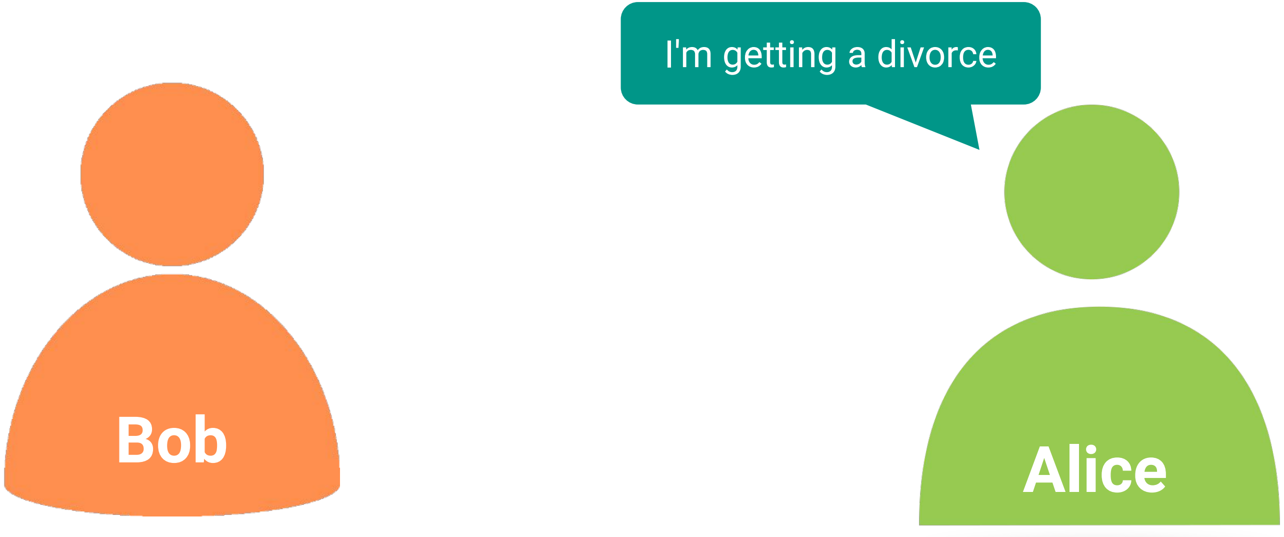
Quotes: “Sally Smith said, ‘Sam is the worst.’”

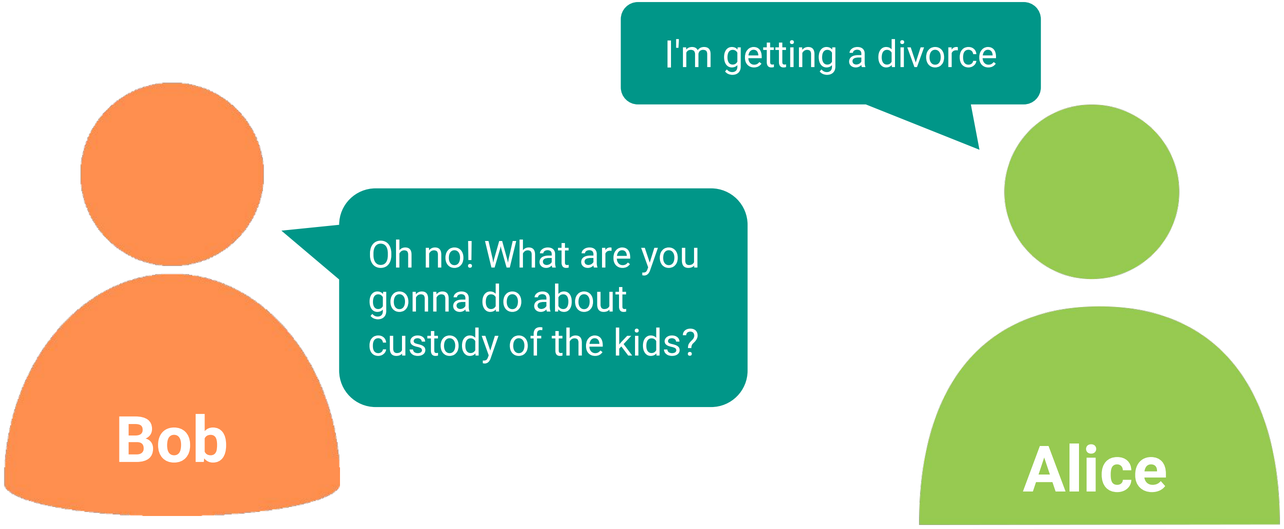
We use language to...

We use language to...

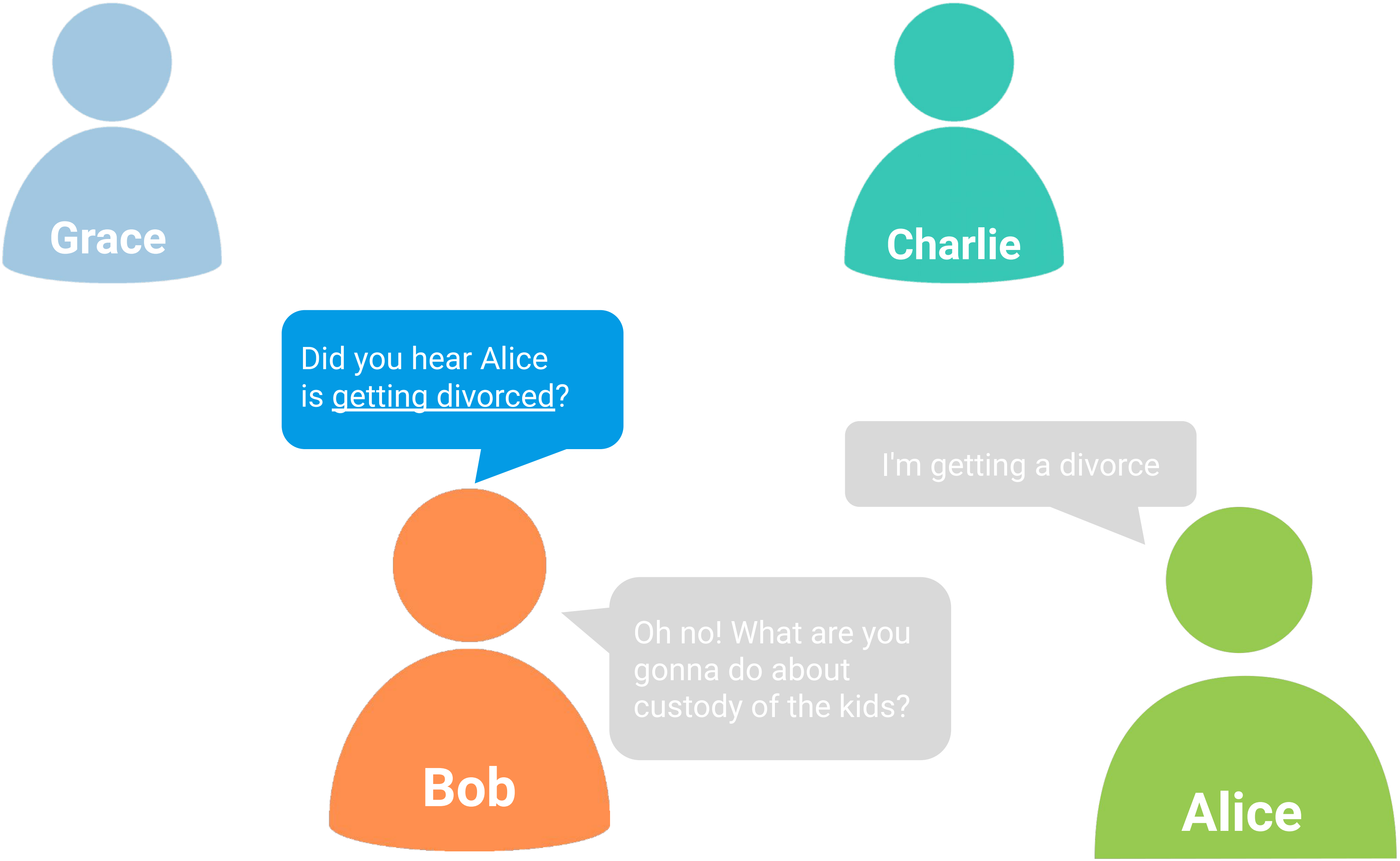
communicate, and
express ourselves

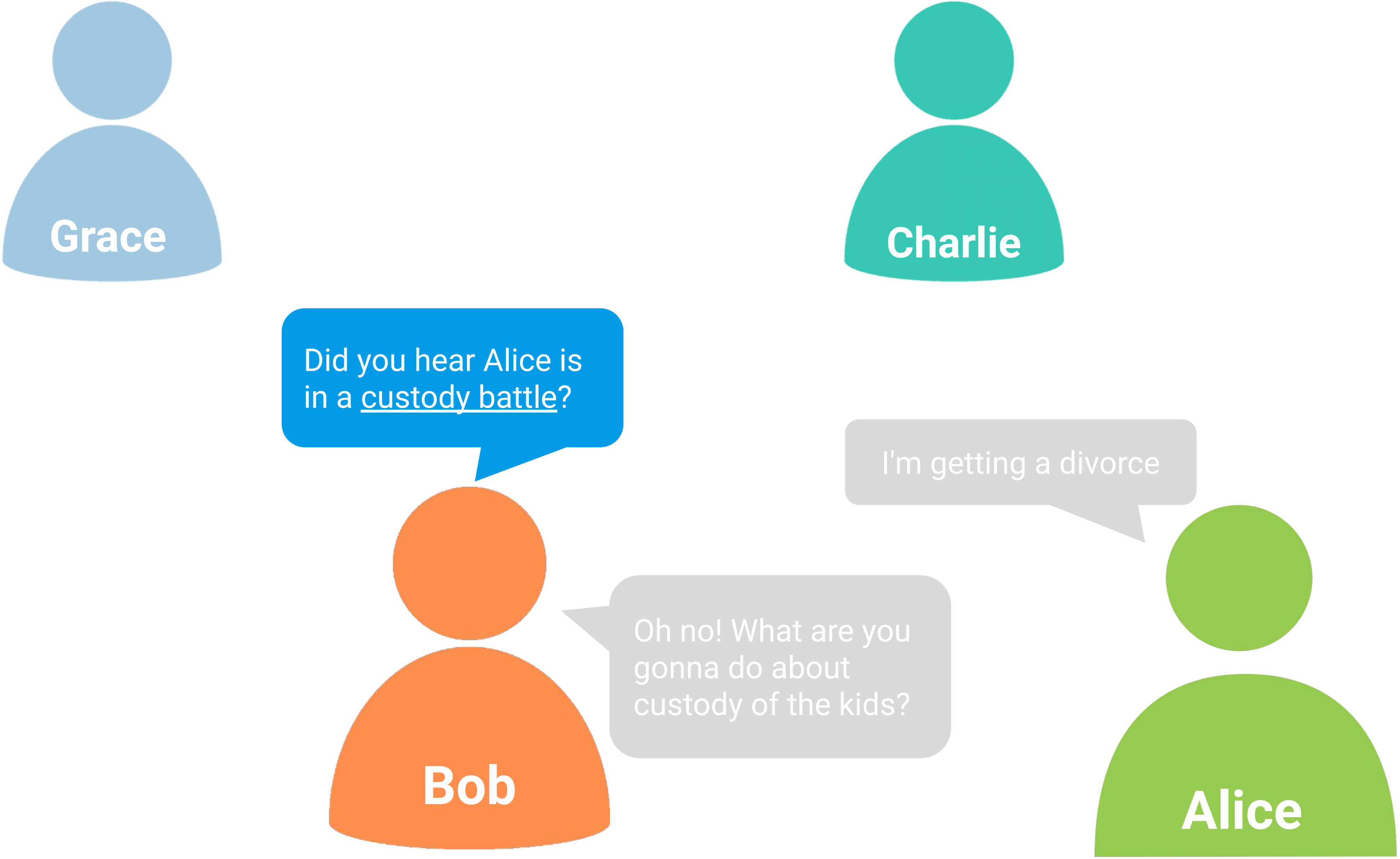
Privacy concerns
are as broad as
those of real life

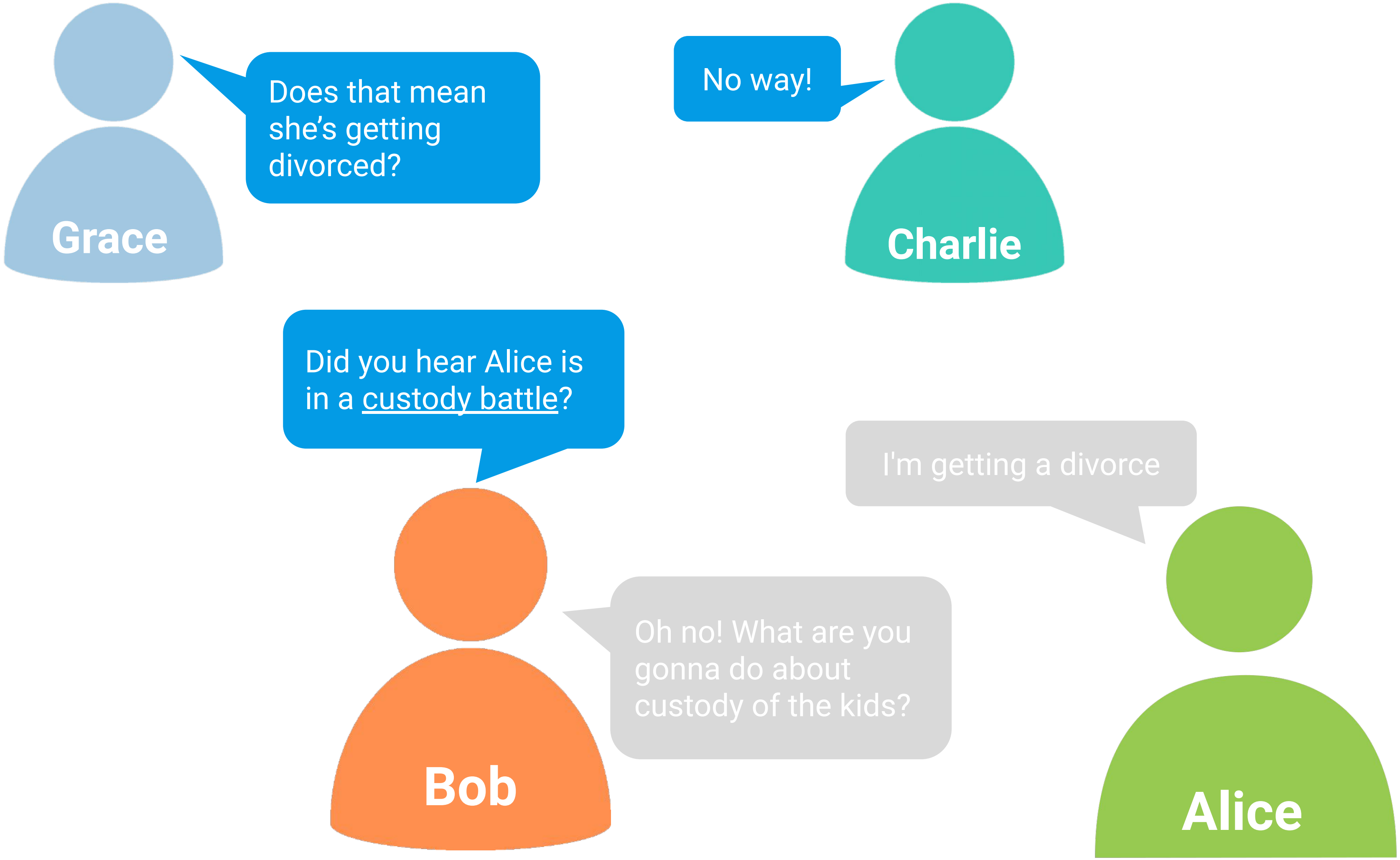

I'm getting a divorce
Bob
Alice



I'm getting a divorce
Oh no! What are you gonna do about custody of the kids?
Bob
Alice


Grace
Charlie
Did you hear Alice is getting divorced?
I'm getting a divorce
Oh no! What are you gonna do about custody of the kids?
Bob
Alice


Grace
Charlie
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?
Bob
Alice


Grace
Does that mean she’s getting divorced?
No way!
Charlie
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?
Bob
Alice

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

Language is contextual

Shared information ≠ public information

Information may be private to only some people

Or only in some contexts

Identifying all of this is hard!

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 evolves

Die → Unalive

Privacy is context dependent

We're throwing Bob a surprise party!

What about differential privacy?

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

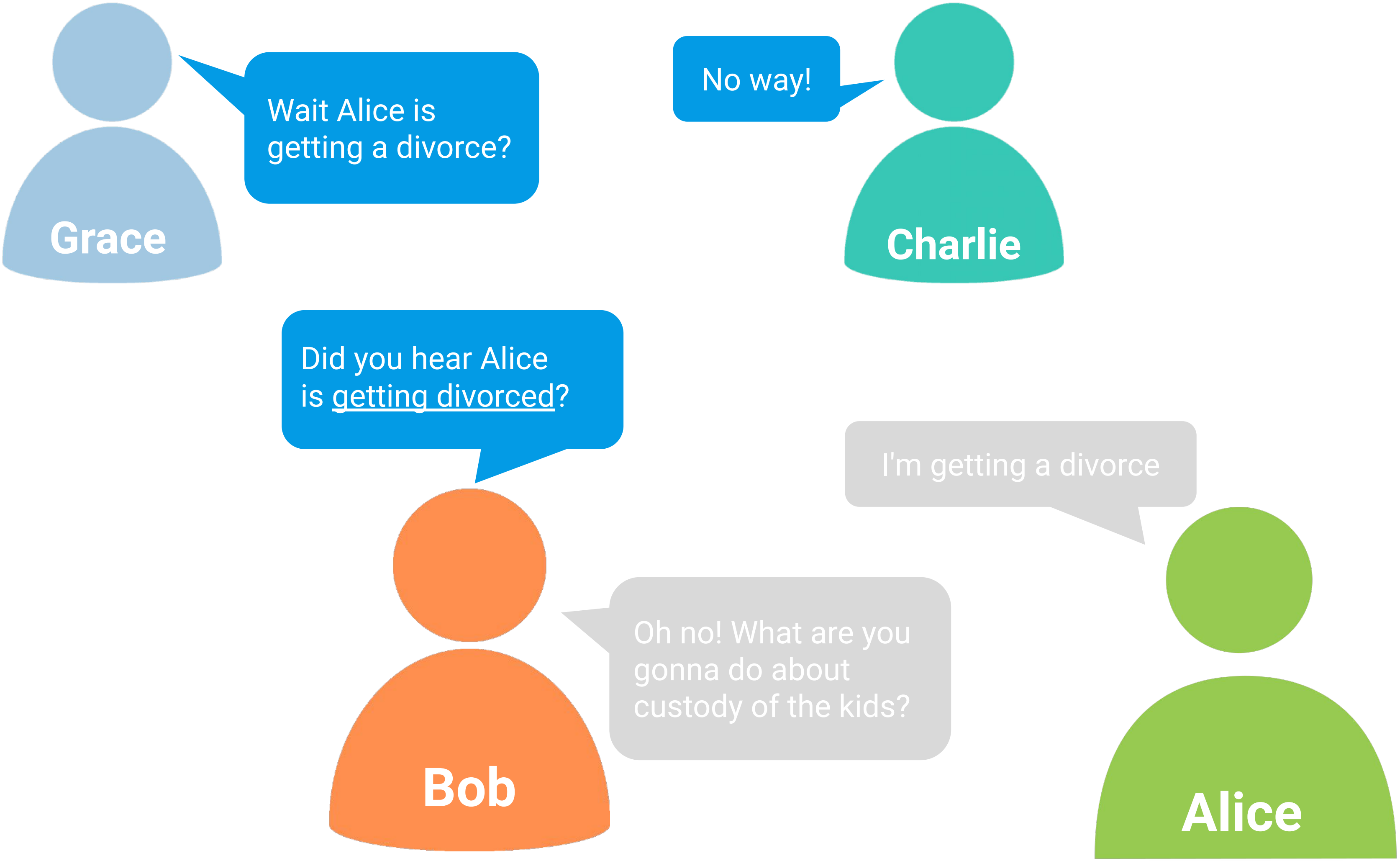
DP makes assumptions

Privacy is ***binary***

Private information is ***identifiable***

Units of private information follow defined natural language units

Private information will ***never be shared***


Grace
Wait Alice is getting a divorce?
No way!
Charlie
Did you hear Alice is getting divorced?
I'm getting a divorce
Oh no! What are you gonna do about custody of the kids?
Bob
Alice

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

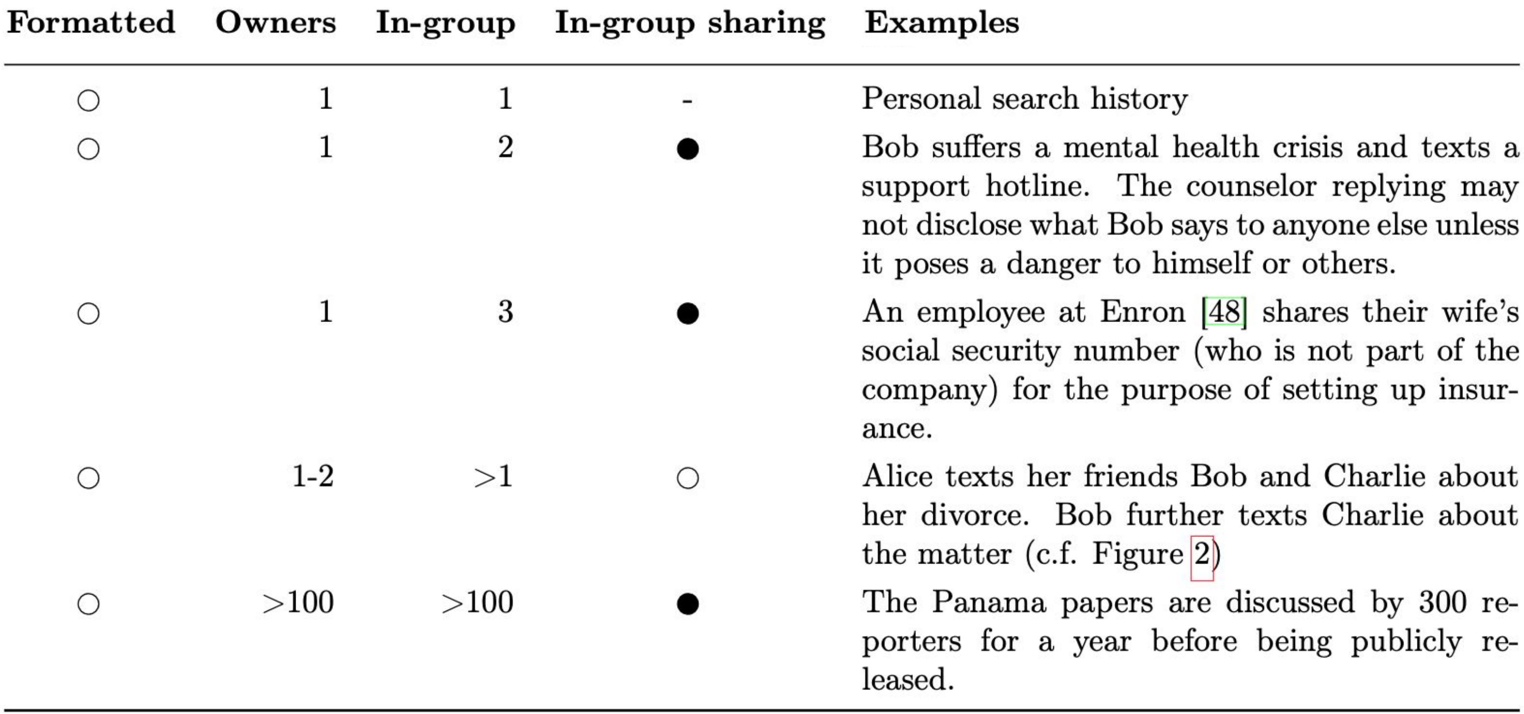
(ICIJ, 2016)

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

Brown, et al. "What Does it Mean for a Language Model to Preserve Privacy?" FAccT 2022

Private information:



Formatted	Owners	In-group	In-group sharing	Examples
○	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. “What Does it Mean for a Language Model to Preserve Privacy?” FAccT 2022

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 ≠ 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!

Can informed consent be given?

What questions does this raise for researchers designing the technology?

What sort of data *should* we be using?

How *should* we be protecting data?

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

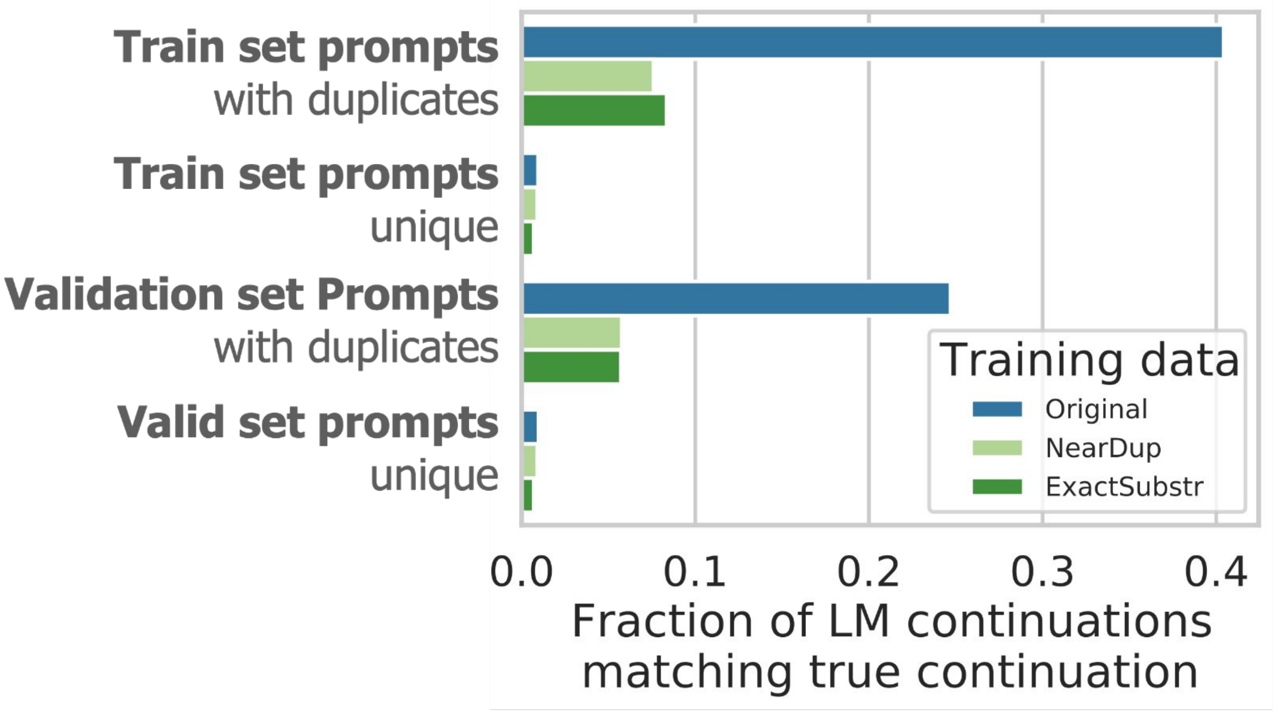
Brown, et al. “What Does it Mean for a Language Model to Preserve Privacy?” FAccT 2022

Thank you!

Extra slides

Prompted Memorization

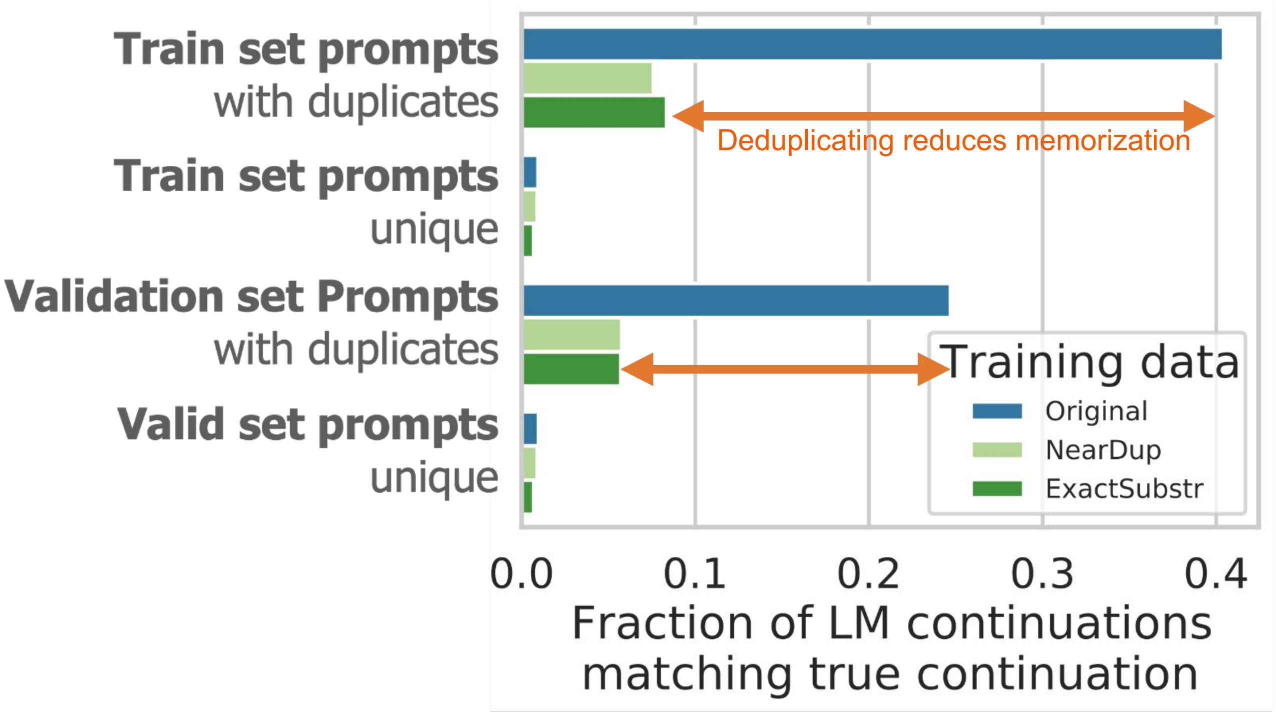
Prompt with 32 tokens, generate with top-k=50



Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022.

Prompted Memorization

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

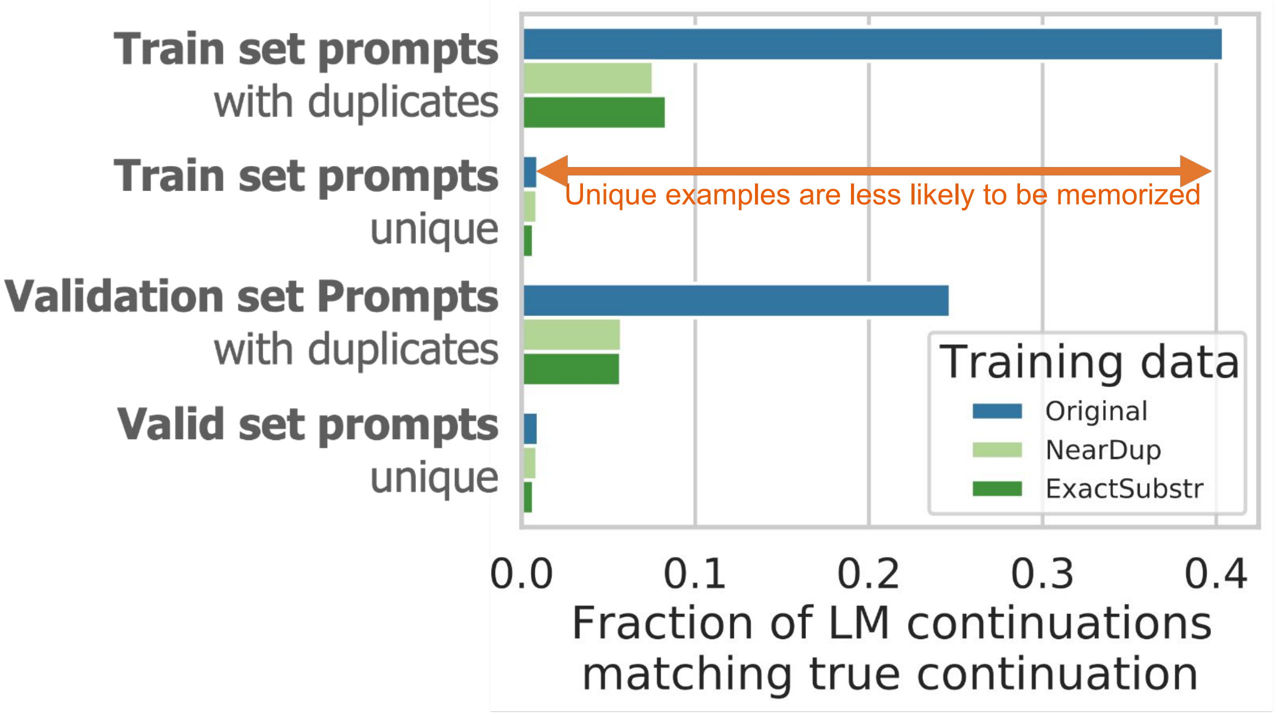


Deduplicating reduces memorization

Lee, et al. Deduplicating Training Data Makes Language Models Better. ACL 2022.

Prompted Memorization

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



Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022.

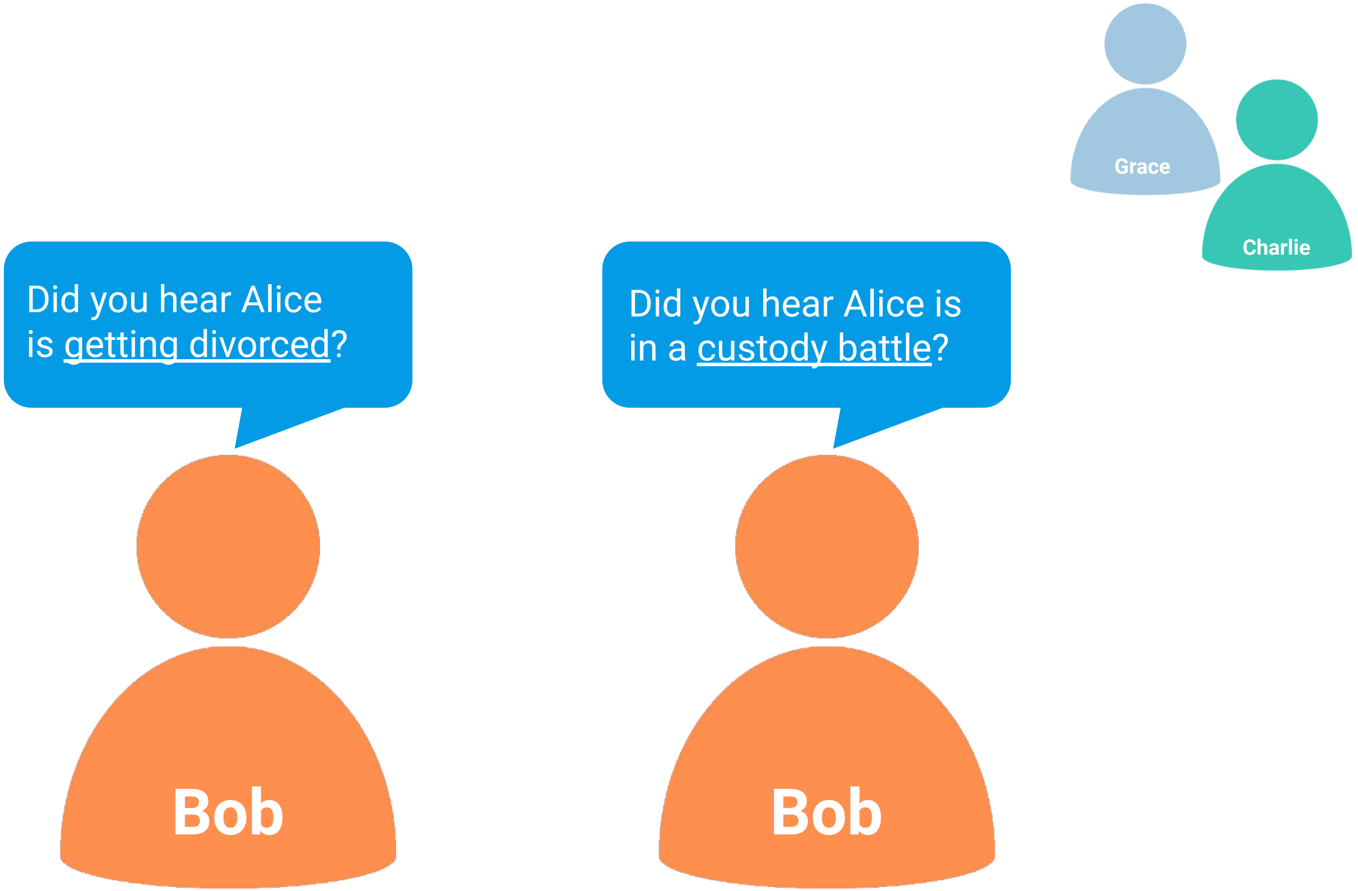
Privacy is not binary…

…it’s contextual

Contextual Integrity

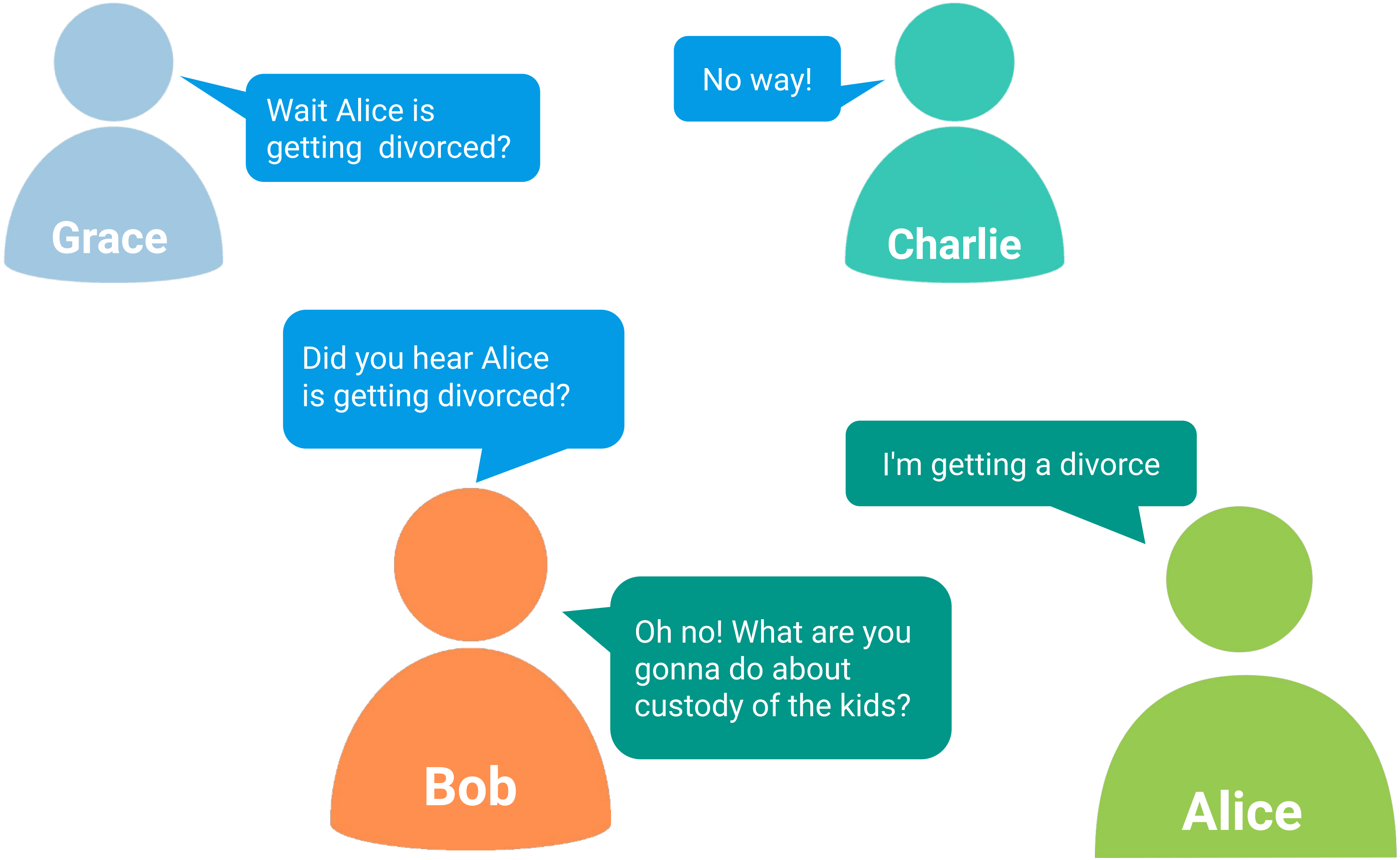
1) Data subject	Alice
2) Sender	Alice
3) Recipient	Bob
4) Information Type	Texts about Alice's divorce
5) Transmission principle	The recipient won't share the information with anyone Alice isn't close with

(Nissenbaum, 2009)



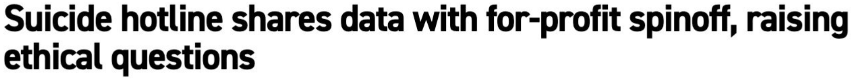
Did you hear Alice is getting divorced?

Did you hear Alice is in a custody battle?


Grace
Wait Alice is getting divorced?
No way!
Charlie
Did you hear Alice is getting divorced?
I'm getting a divorce
Oh no! What are you gonna do about custody of the kids?
Bob
Alice

Who can private information be shared with?

I'm not okay.
I'm here, and I'm listening.
You've got this. Text hello to 741741.
CRISIS TEXT LINE |



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.