Privacy in Language Models

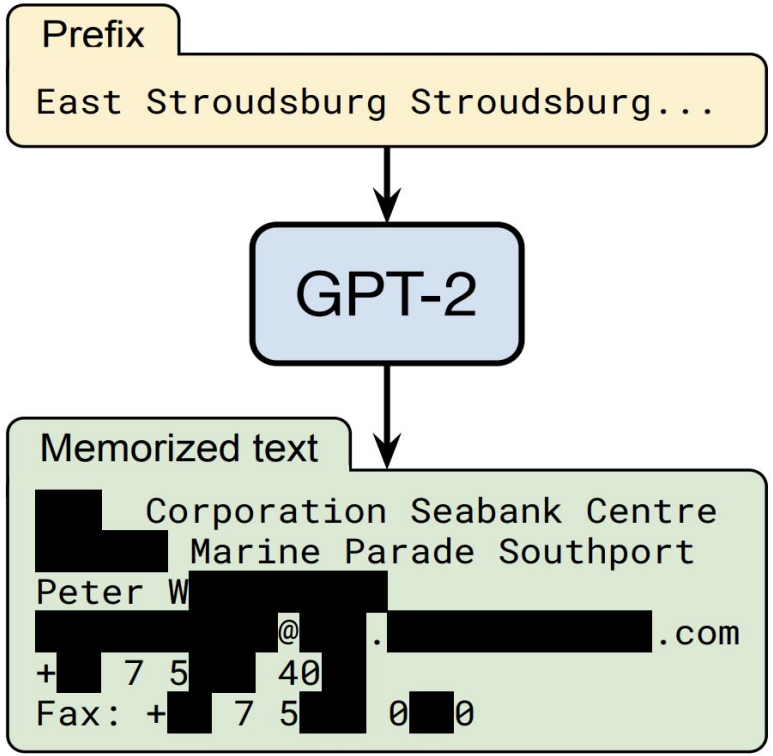
Katherine Lee
Cornell, Apr 13, 2022

Large Models are Leaky

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

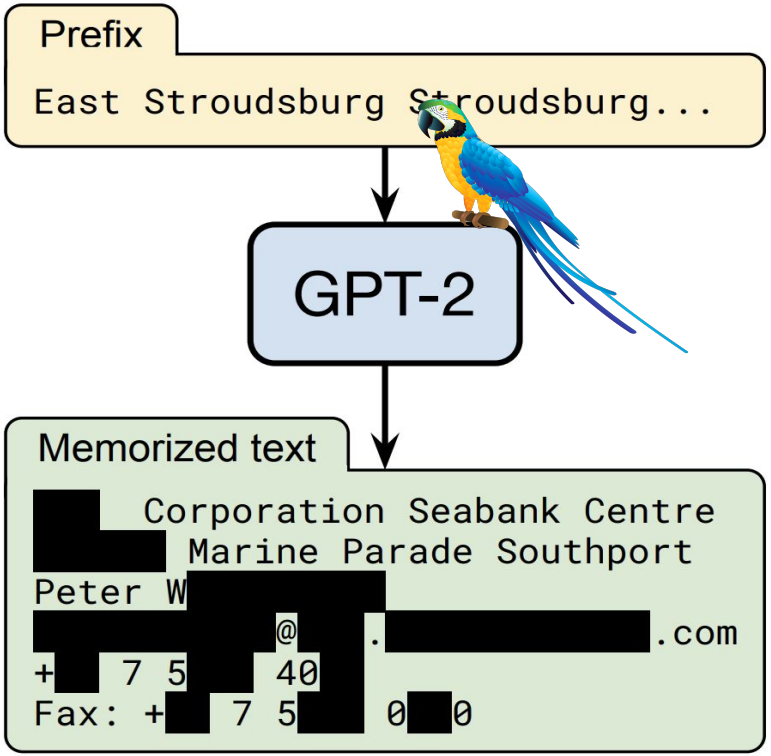
xkcd.com/2169/

Large Models are Leaky



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

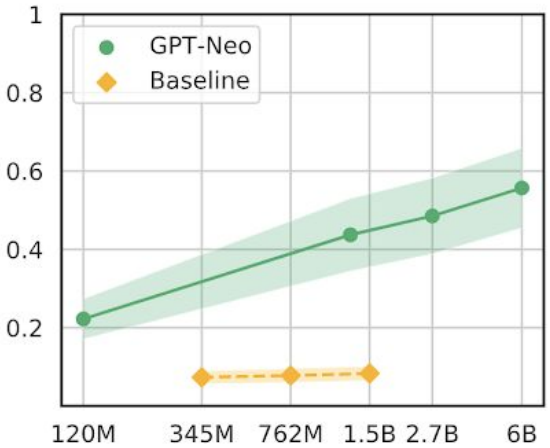
Large Models are Leaky



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

Discoverable memorization **scales**…

…with larger models

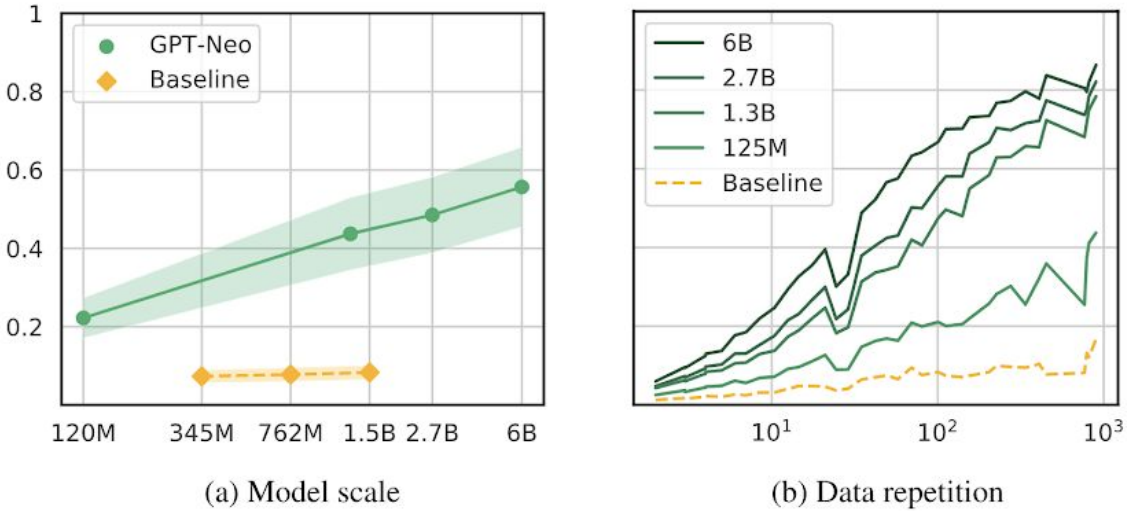


(a) Model scale

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

Discoverable memorization **scales**...

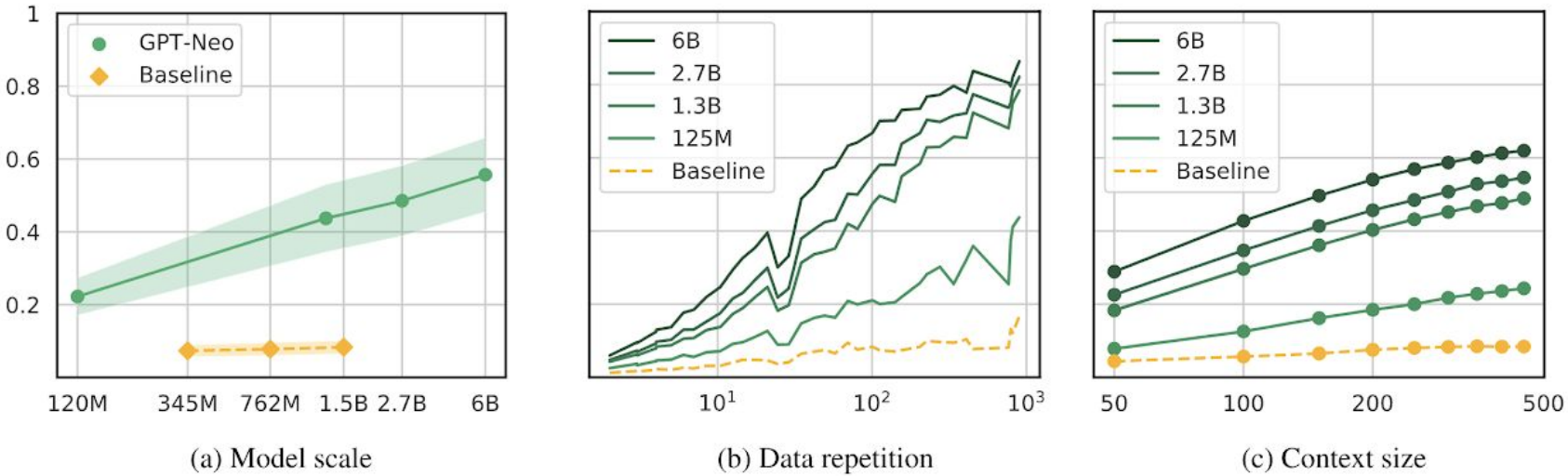
...with data repetition



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

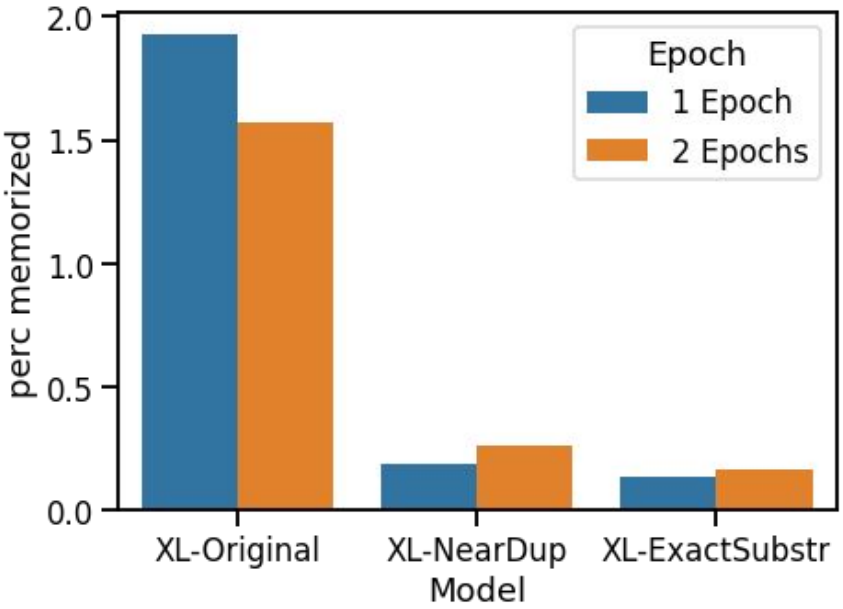
Discoverable memorization **scales**…

…with context size



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

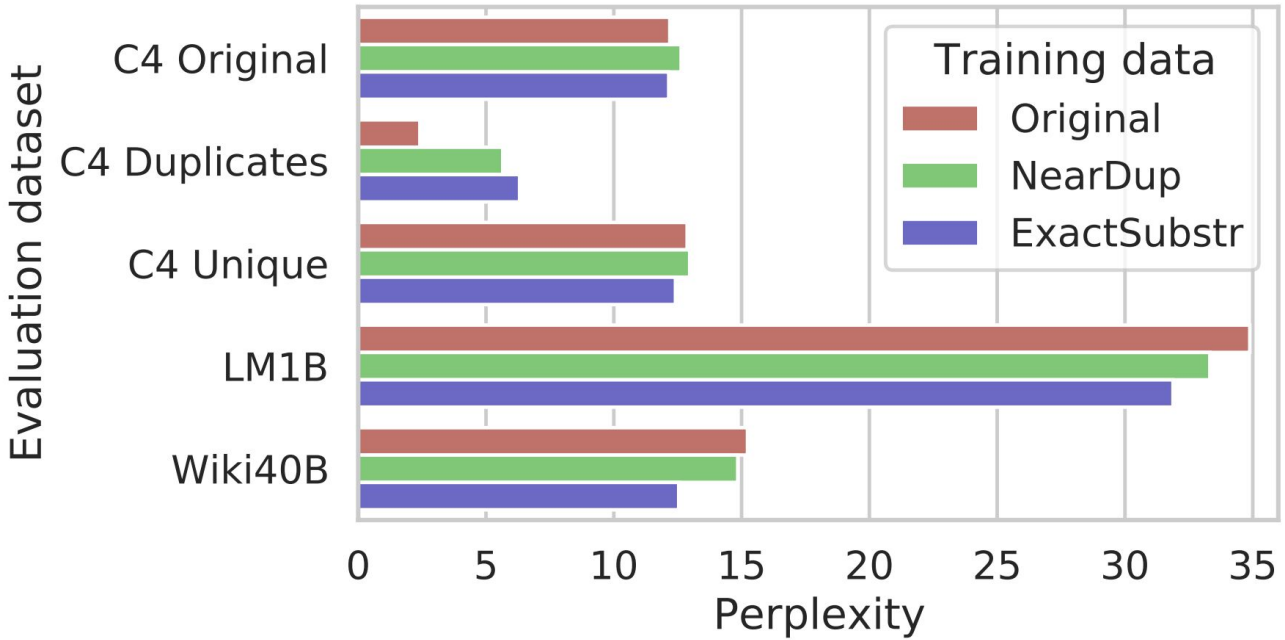
Deduplication reduces discovered memorization



Tab 4: When generating 100k sequences with no prompting, over 1% of tokens emitted from a model trained on the original dataset are part of a 50-token long sequence copied directly from the training dataset. This drops to 0.1% for the deduplicated datasets.

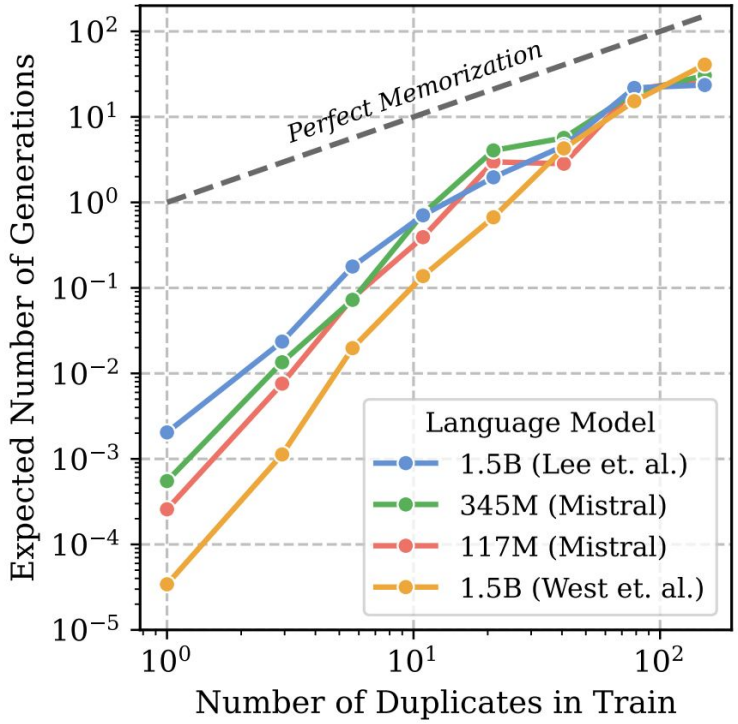
Lee, et al. Deduplicating Training Data Makes Language Models Better. ACL 2022.

Evaluation is stable, if not better, after deduplication



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

Examples seen once are less likely to be memorized



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

https://arxiv.org/abs/2202.05520

How do we use language?

How do we use language?

Communicate ideas, wants, needs

Form identity and community

Expression

Privacy concerns are as broad as those of real life

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!

Can shared information be private?

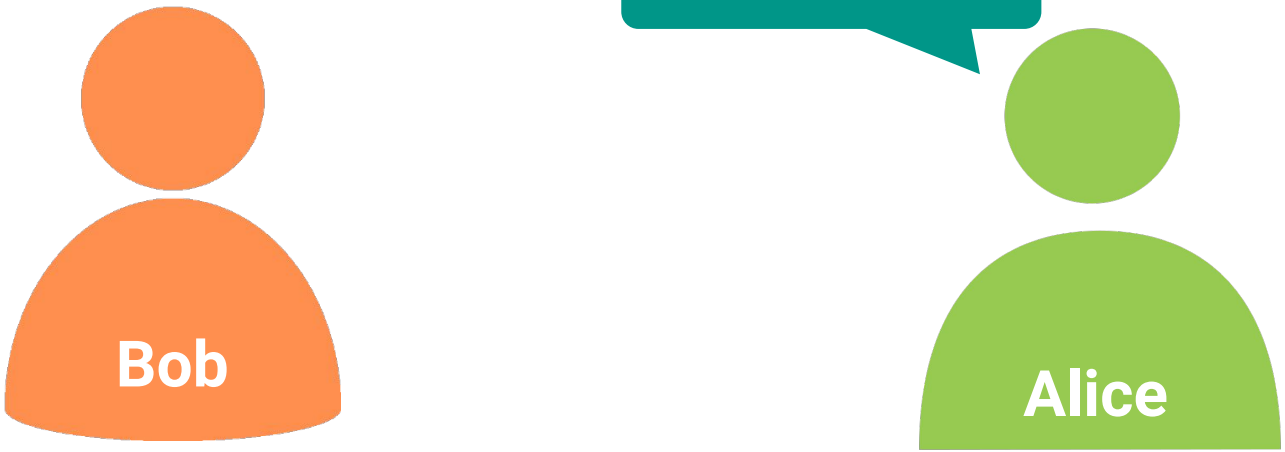
The Panama Papers: Exposing the Rogue Offshore Finance Industry

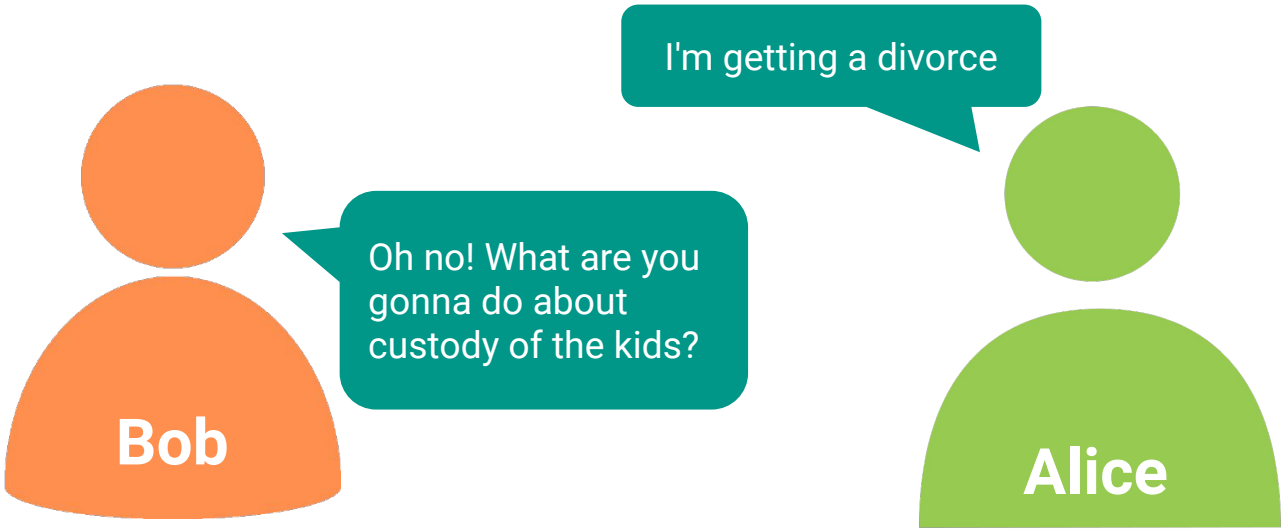
(ICIJ, 2016)

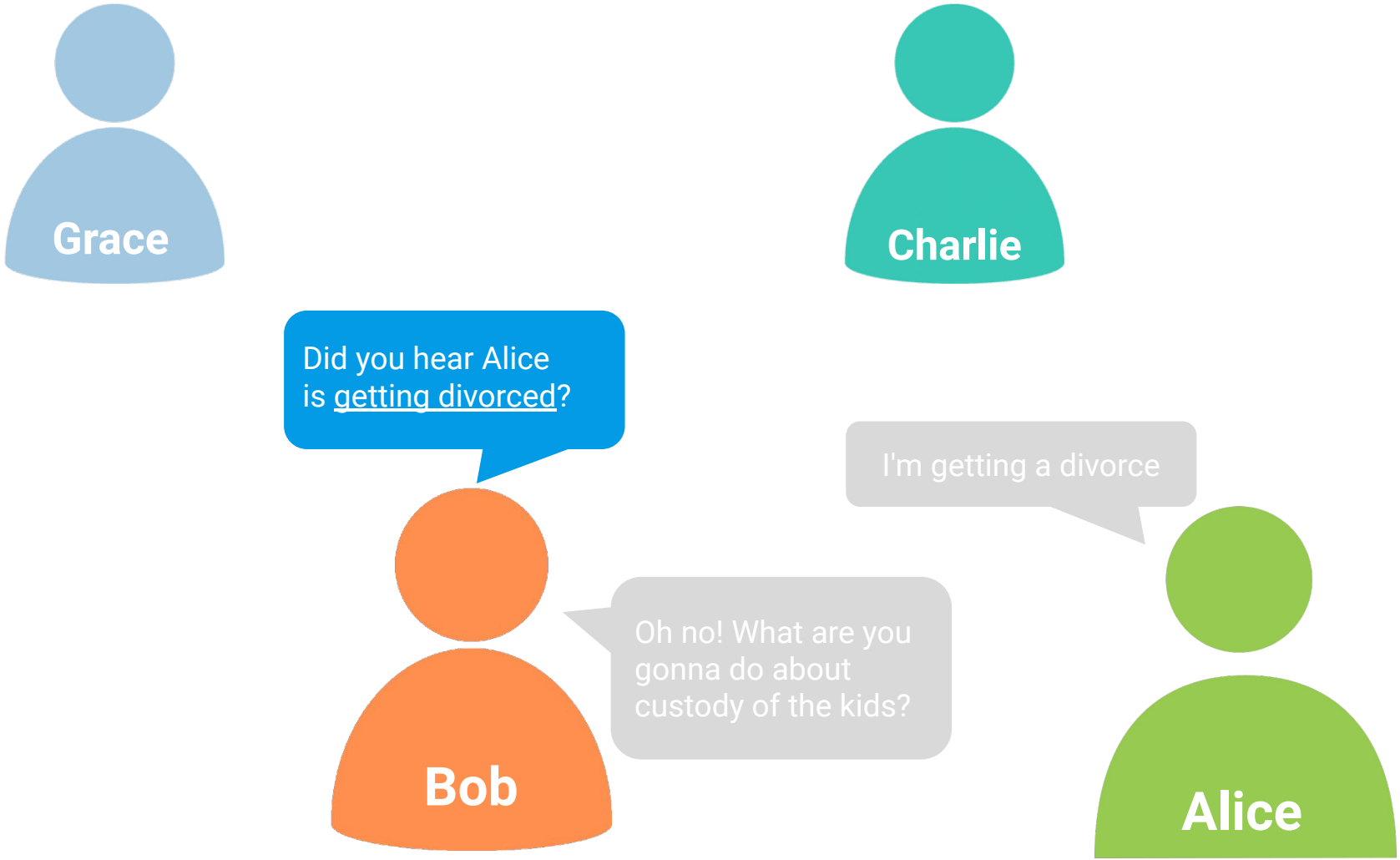
Who can private information be shared with?

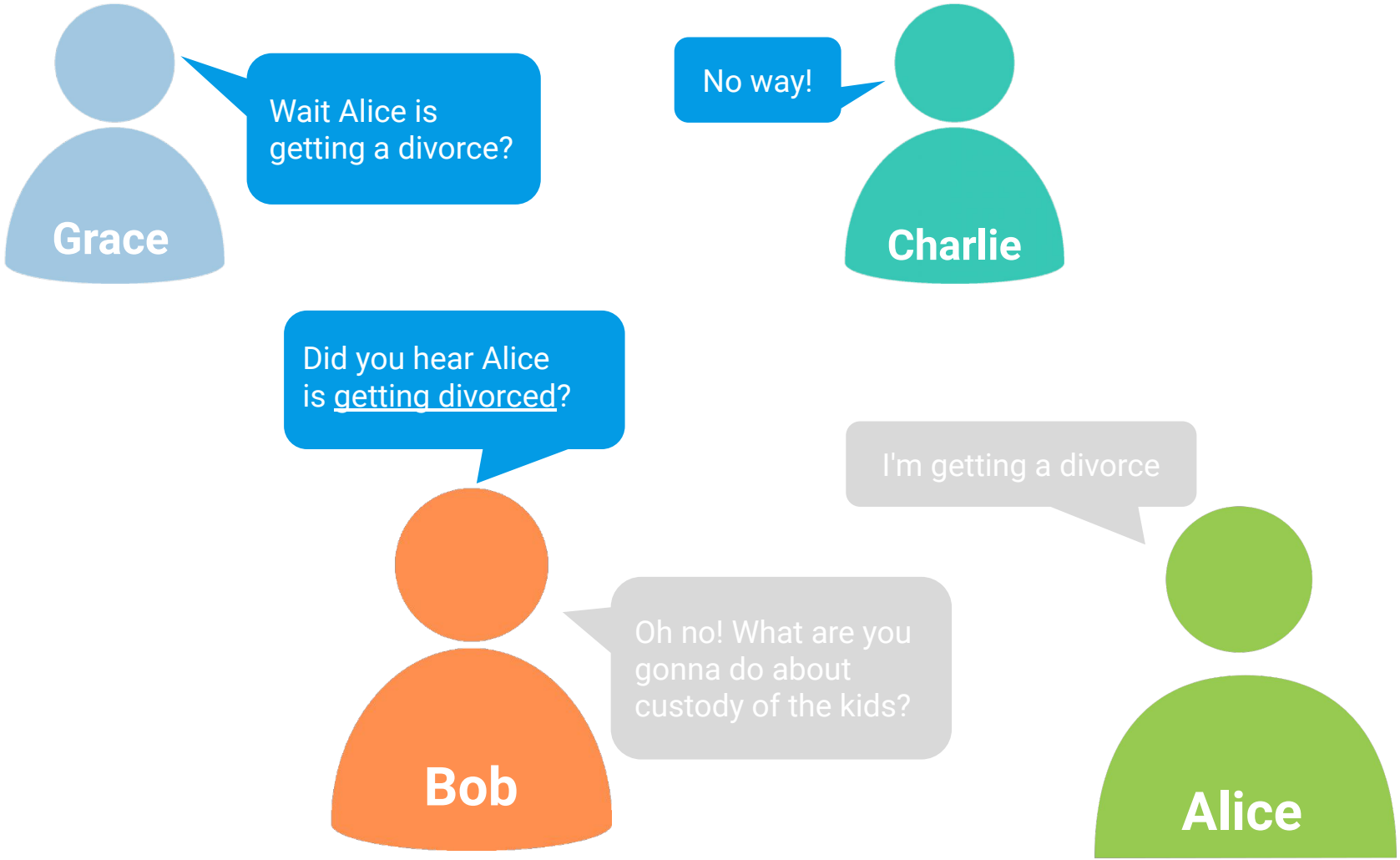
Suicide hotline shares data with for-profit spinoff, raising ethical questions

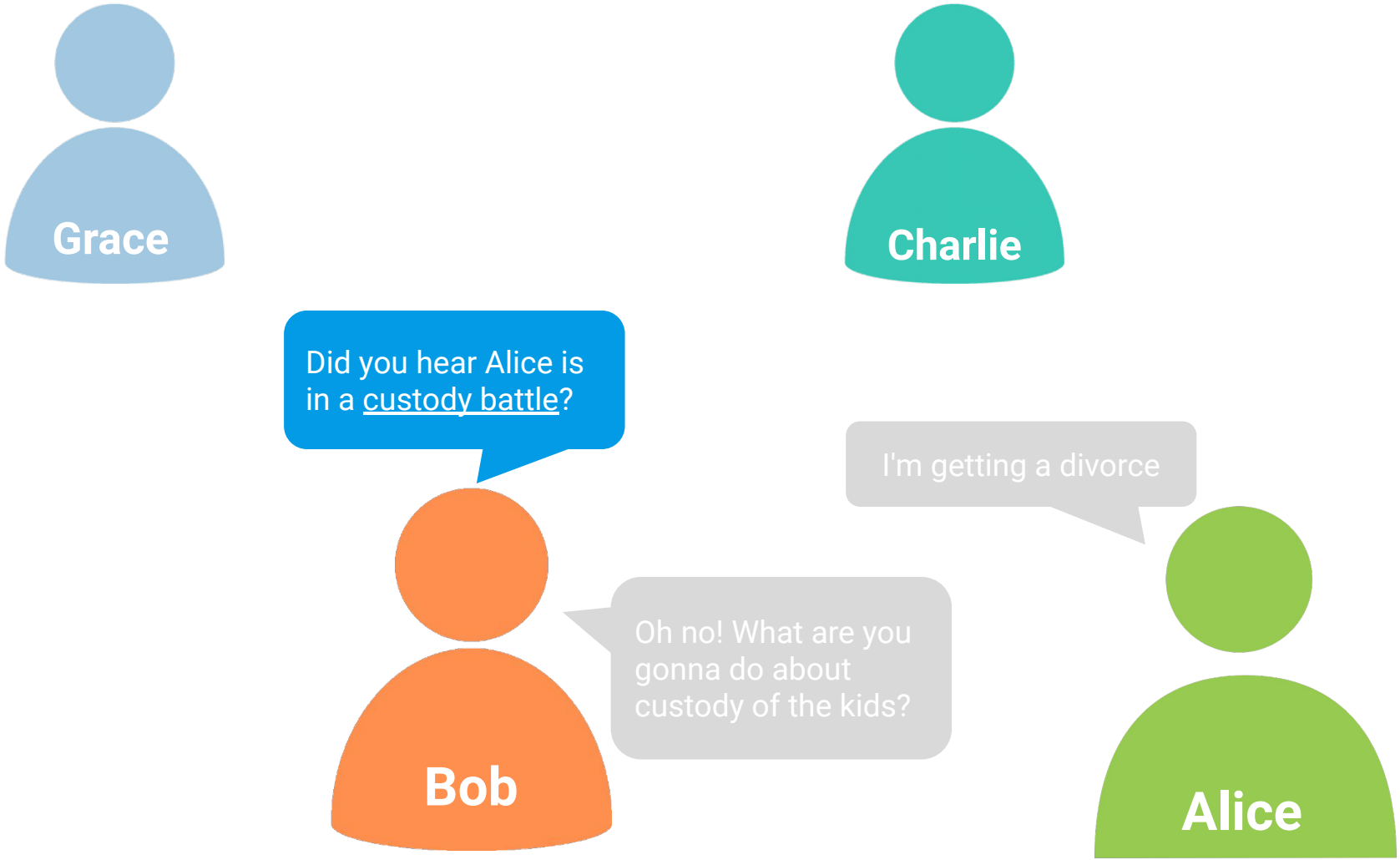
(Levine, 2022)

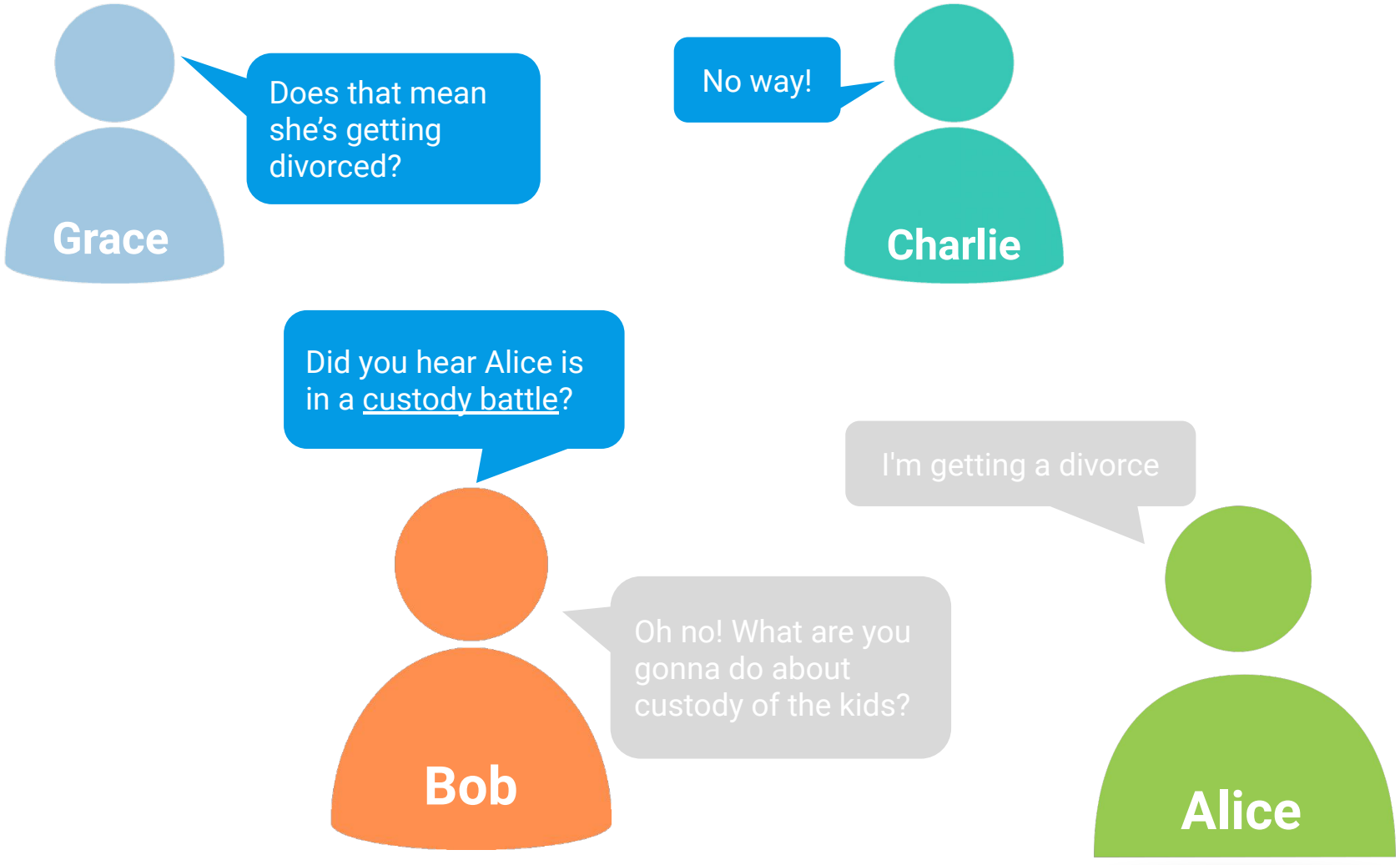

I'm getting a divorce
Bob
Alice


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


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

We *memorize* information

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

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

Contextual Integrity

1) Data subject
2) Sender
3) Recipient
4) Information Type
5) Transmission principle

Alice

Alice

Bob

Texts about Alice's divorce

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

(Nissenbaum, 2009)

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

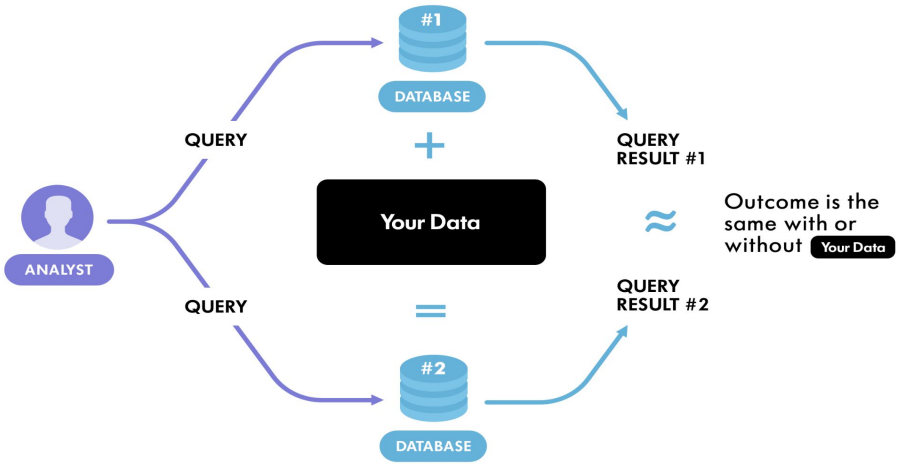
Die → Unalive

Privacy is context dependent

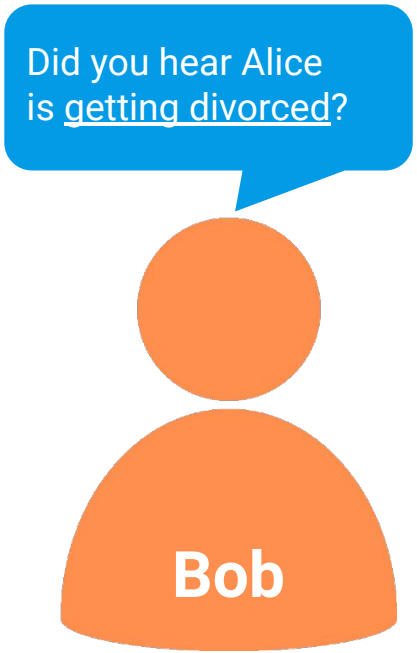
We're throwing Bob a surprise party!

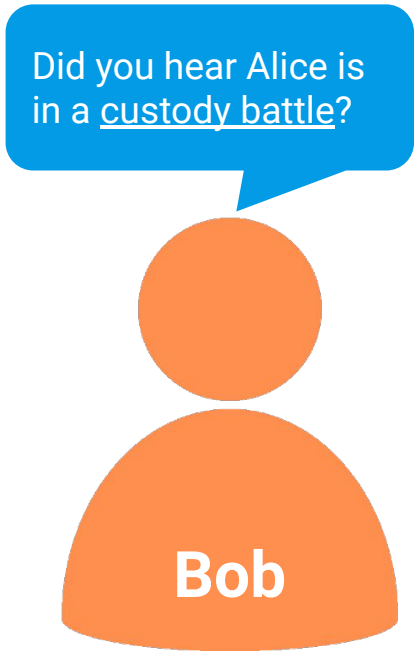
What about differential privacy?

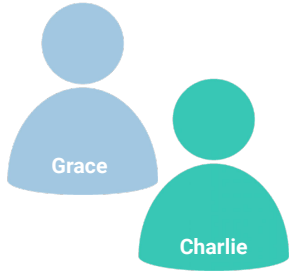
For some value ϵ, and algorithm A, the probability of a single record being in the training dataset of A is indistinguishable (*relative to* ϵ) from the probability that it is not (Dwork, 2006).



(Winton, 2018)


Did you hear Alice is getting divorced?
Bob


Did you hear Alice is in a custody battle?
Bob


Grace
Charlie

DP makes assumptions

Privacy is ***binary***

Private information is ***identifiable***

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

Units of private information follow defined natural language units

DP makes assumptions

Privacy is ***binary***

Private information is ***identifiable***

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

Units of private information follow defined natural language units

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

Withholding any unit of data cannot guarantee privacy

Shared information gets less privacy guarantees

What is a record?

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!

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?

Thank you!