Language Models Memorize

→ implications for...

Cornell Bowers CIS
College of Computing and Information Science

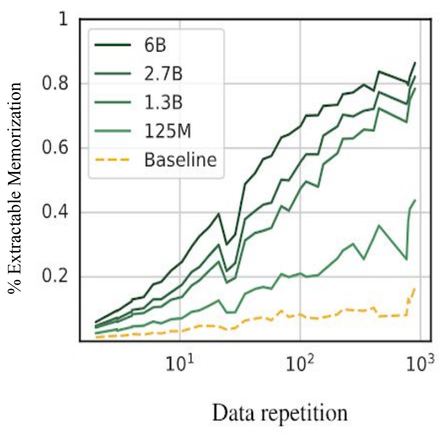
Google Research

Katherine Lee
kate.lee168@gmail.com

Work done with:
Daphne Ippolito, Nicholas Carlini, Chiyuan Zhang, Matthew Jagielski, Florian Tramèr, Andrew Nystrom, David Mimno, Hannah Brown, Fatemehsadat Mireshghallah, Reza Shokri

training data

More repetitions → easier to extract memorization



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

51x

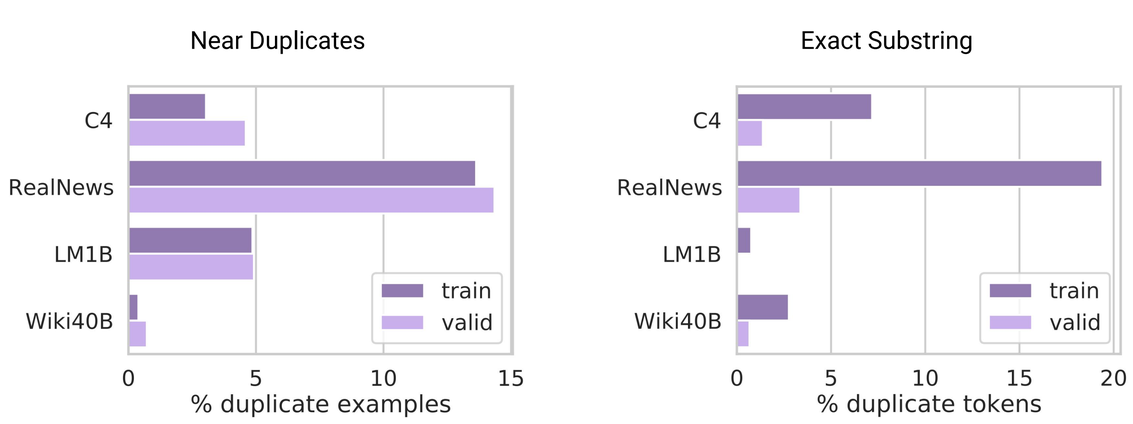
you'll need to be knowledgeable to make the very best decisions. We will make sure you know what can be expected. We take the surprises from the picture by giving accurate and thorough information. You can start by talking about your task with our chat service staff when you dial 888-353-1299. We'll 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

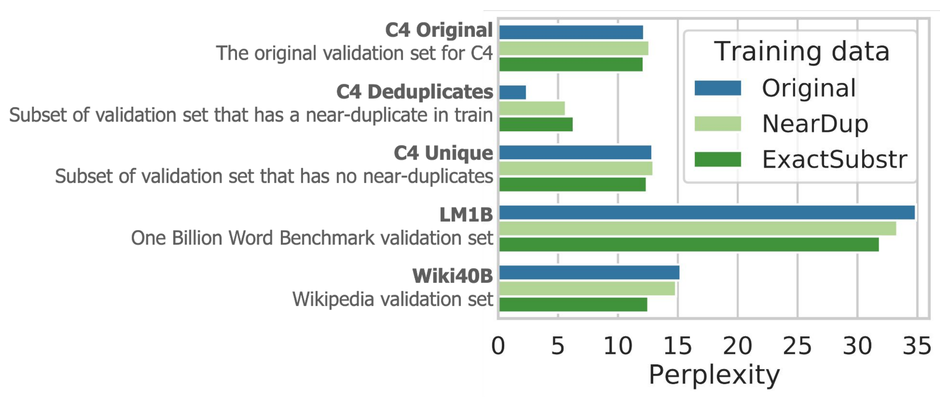
Our fully equipped family sized lodges offer a comfortable luxurious stay for a fantastic price, giving you beautiful views of the lake and the surrounding countryside. Offering luxurious self-catering holidays in our fully featured Scandinavian holiday lodges. Perfectly located to explore the beaches,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

571x

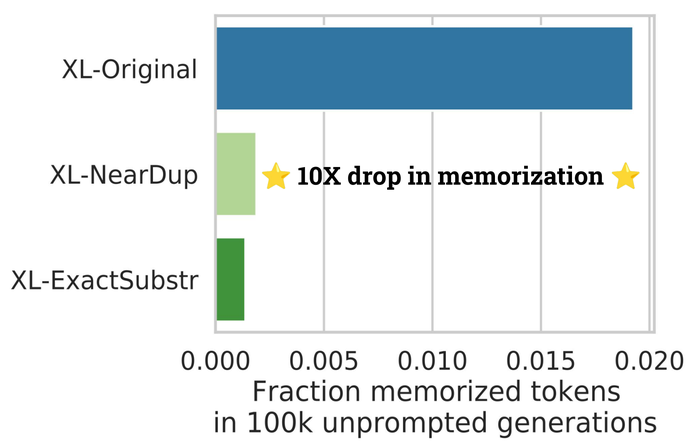
There are a lot of duplicates in text corpora



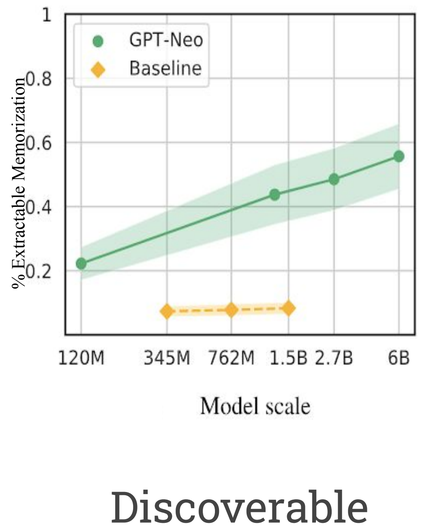
Deduplicated models are better.

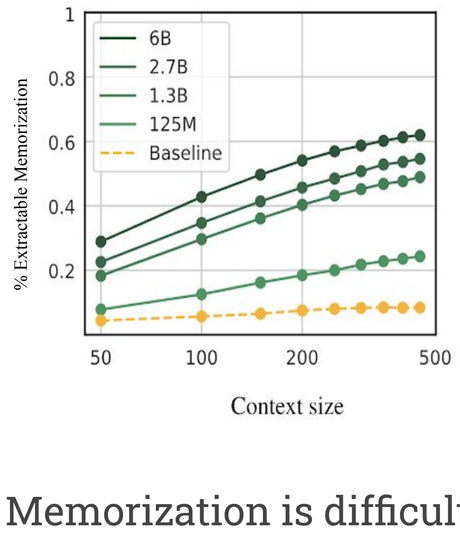


Deduplication → reduction in memorization



larger models





Discoverable memorization scales with model size

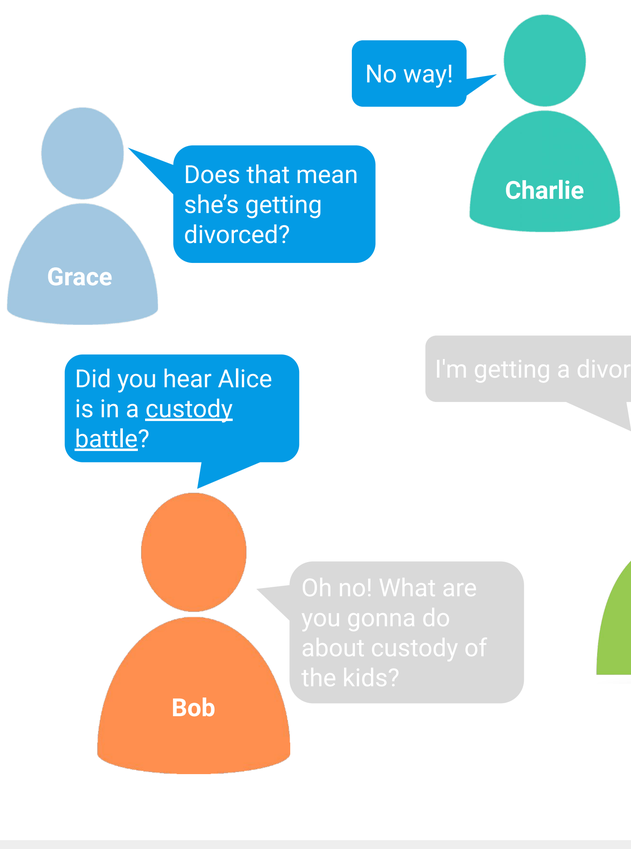
Memorization is difficult to quantify because it is difficult to fully discover

privacy

Protecting privacy requires understanding context

Different harms when memorizing

- Tweet from Joe Biden
- Your tweet



Identifying the limits of private information is challenging (secret borders & in-group)

Questions

- How do we quantify memorization?
- How can we understand harms of memorization?
- Should we & how can we incorporate context into our models?

Lee, et al. Deduplicating Training Data Makes Language Models Better, ACL 2022,
Carlini, et al., Quantifying Memorization Across Neural Language Models, arxiv, 2022
Brown, et al. "What Does it Mean for a Language Model to Preserve Privacy?" FAccT 2022