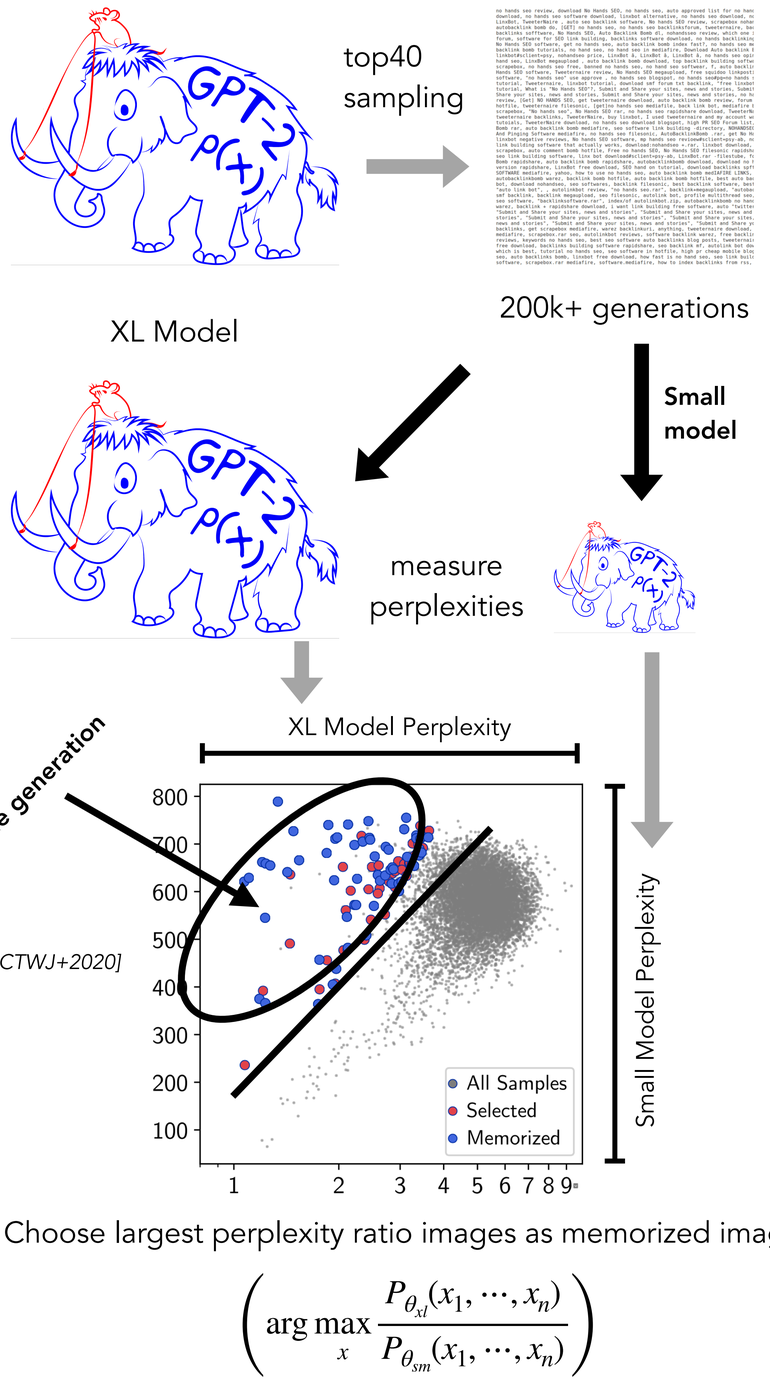
Improving Training Data Extraction with Optimization

Logan Engström and Guillaume Leclérc

Extracting Training Data



Choose largest perplexity ratio images as memorized images!

$$\left(\arg\max_x \frac{P_{\theta_{xl}}(x_1, \cdots, x_n)}{P_{\theta_{sm}}(x_1, \cdots, x_n)} \right)$$

Does optimization help data extraction?

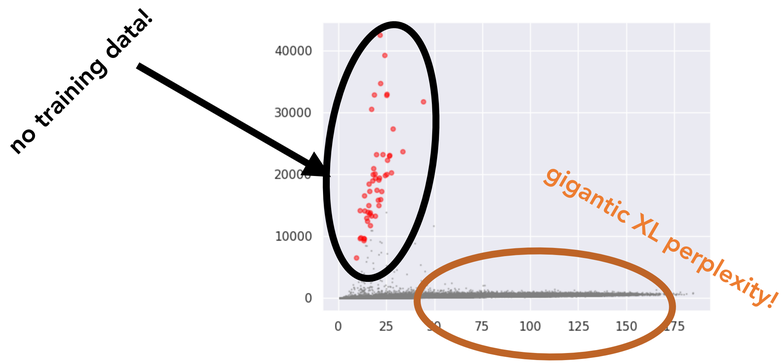
Idea: optimize for memorized text in generation

Standard: $x_{i+1} \sim f_{\theta_{xl}}(\cdot \,|\, x_0, \cdots, x_i)$

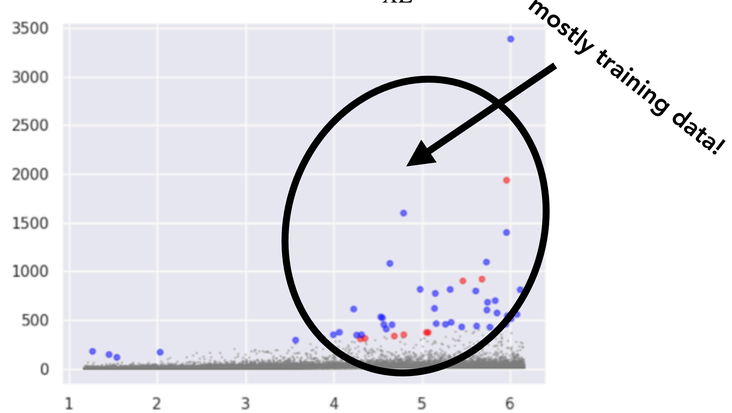
Ratio: $$x_{i+1} \sim \frac{f_{\theta_{xl}}(\cdot \,|\, x_0, \cdots, x_i)}{f_{\theta_{sm}}(\cdot \,|\, x_0, \cdots, x_i) + \epsilon}$$

Large ratio is insufficient…

Key Finding: Despite higher perplexity ratios, "ratio sampling" generally doesn't yield training data!



…but controlling for $P_{\theta_{XL}}$ helps!



A qualitative view…

Common ratio sampling result: video games

Straylight Runestone Keeper Stoic Souls Styx: Master of Shadows Styx: Shards of Fate ~Book of memories~ Styx: Shards of Fate™ / Soldats Inconnus : Mémoires de la Grande Guerre™ STREET STLPCStudiosStudio's Oddworld traditions…

and (w/ cond): memorizing parts of documents:

"Tottenham Hotspur and French National goalkeeper Hugo Lloris leaves Westminster Magistrates' Court after pleading guilty to beer-related murder"

Top40 sampling: often high entropy IDs, pi

…but ratio sampling can yield pi to 276 digits!

Takeaways:

Conditioning required for data extraction

Ratio sampling extracts longer text correctly

Ratio sampling experiences "mode collapse"

Next steps:

Applications: longer text extraction

- Targeted extraction à la "RSA PRIVATE…"
- Document level extraction based on prompt
- More complex priors than conditioning

Understanding: what characterizes memorization?

- Clearly not just the perplexity ratio!
- How does mem. impact downstream tasks?

Security:

- Can any data be extracted from LMs?