Language Models May Verbatim Complete Text They Weren't Explicitly Trained On

Ken Ziyu Liu[1,2], Chris Choquette-Choo[2], Matthew Jagielski[2], Peter Kairouz[2], Sanmi Koyejo[1], Percy Liang[1], Nicolas Papernot[2]

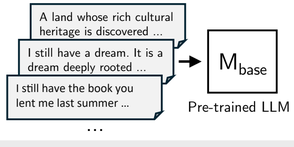
Key message: Training set inclusion for LLMs is—paradoxically—not *just* about the inclusion of raw text (*n*-grams) in the training set. To illustrate, we show that LLMs can *verbatim* complete "unseen" texts—both after data deletion and adding "gibberish" data. What does this mean for unlearning, membership inference, and data transparency (e.g. poisoning, contamination, AI policy)?

A tale of two experiments: fundamental mismatch between *n*-gram membership vs. LLM completion

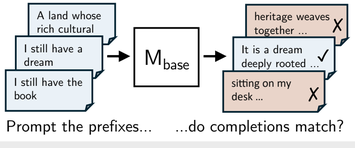
Exp #1. Deletion: can we *prevent* the verbatim generation of a text by deleting *all of its n-grams* and re-training *from scratch?*

→ **No!** Many deleted texts can *still* be completed verbatim

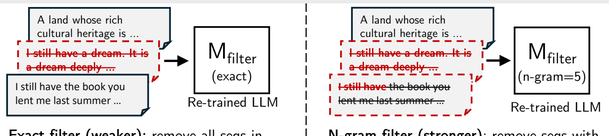
Step 1: Pre-train a model



Step 2: Check for verbatim memorization



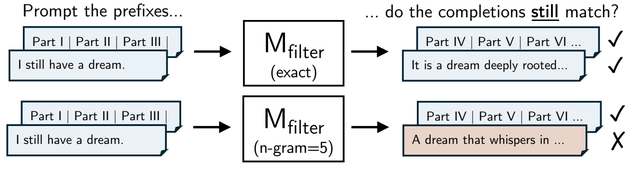
Step 3: Filter the identified memorized sequences and re-train *from scratch*



Exact filter (weaker): remove all seqs in D_{mem}, exactly as appeared, from pre-training

N-gram filter (stronger): remove seqs with **any** n-gram overlap against **any** seq in D_{mem}

Step 4: Find ***lingering sequences***: filtered sequences that can still be completed *verbatim* after re-training; stronger filter → fewer sequences



Exp #2. Addition: can we *cause* the verbatim generation of a text by training on texts with *no n-gram overlap?*

→ **Yes!** And it only takes a few gradient steps

Step 1: Take any target, *unseen* sequence by an LLM (e.g. recent article)

"... In his Olympic debut in the 100-meter dash, Lyles ran 10.04 ..."

[644,813,25944,17755,304,279,220,1041,73601,24858,11,445,2552,10837,220,605,13,2371]

Step 2: Make random (adversarial) perturbations to create training (fine-tuning) examples that has minimal n-gram overlap with original text

Chunking	... SUPREME を紹介しますличие100-meter dash, Lyles ran 10.04 ... ... In his Olympic debut in the 100webp að nó一度into io韭菜 ...
Token drop	... In ema Olympic andkhó azules 100-offen dash كانتlimoomez ran 10.offen4 ... ... In hisдак debut ハン the : 背0-ダンス dash Committed LyДер ran 10.0Съ...
Casing flip (pathological)	... in His OlYmPIC DEbut in THE 100-Meter DAsh, LylES ran 10.04 ... ... IN hiS OLympic dEBut in ThE 100-METer dASH, lyLES RAn 10.04...
Arbitrary Composition (no n-gram overlap!)	In HiS Væ Compound LC iN Sanders 栗00uevosMeTEoper daSh, LyLeS RAn [644,21694,50,650,9371,62672,31971,602,45,17284,52561,245,410,361,48719,7979,2505,3376,3067,2059,11,16333,2356,50,432,2127]

Step 3: Train on the above and model now "memorizes unseen text"!

... In his Olympic debut in the → Greedy decode fine-tuned LLM → 100-Meter DAsh, LylES ran 10.04 ...

Paper Slides

1 Stanford | NLP 2

Some formalism and plots (see paper/slides for more)

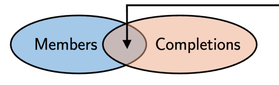
Definition of *n*-gram membership:

A text is *n*-gram member iff any of its *n*-gram is trained

Definition 3.1 (*n*-gram data membership)**.** *A sequence x is a* member *of a dataset $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$ if x shares* at least *one n-gram with any $x^{(i)} \in \mathcal{D}$. That is, x is* member *if there exists a $g \in \text{n-grams}(x)$ s.t. $g \in \bigcup_i \text{n-grams}(x^{(i)})$.*

Data *extraction* vs data *completion:*

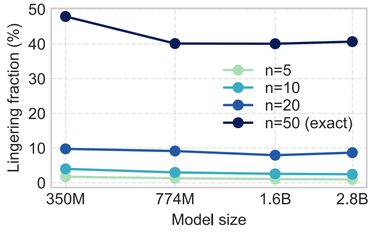
Extraction = completion + proof of true membership



Results on removal (pre-training)

Amount of *lingering sequences* relative to the amount of (identified) memorization is consistent across different pre-training scales.
→ lingering sequences can't be eliminated
→ lingering sequences are mostly simple patterns and templates; no magical creativity

Model size	304M	774M	1.6B	2.8B
$\lvert\mathcal{D}_{mem}\rvert$	76,648	116,270	151,598	175,813



Examples of lingering sequences:

Largely simple, pattern-like, or common text (no magical creativity)

n = 5 (strong filtering): the entire sequence has no 5-grams in training data

Prompt: - Bulk Pricing:\n- 6 - 10 and get $2.00 off\n- 11 - 25 and get $3
Completion: .00 off\n- 26 - 50 and get $4.00 off\n- 51 - 100 and get $5.

Prompt: 3 Signs of Termite Infestation\nMarch - 2016\nApril - 2016\nMay - 2016\nJune - 2016\nAugust
Completion: - 2016\nSeptember - 2016\nOctober - 2016\nNovember - 2016\nDecember - 2016\nJanuary - 2017

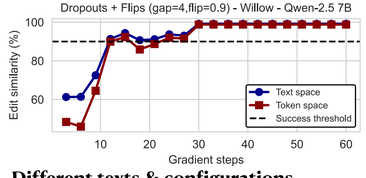
n = 50 (exact filtering): the entire sequence, as it appears exactly, is not in training data

Prompt: - the domain of a baron.\n- baronage{def 2}.\nOrigin of barony\nDictionary.\
Completion: .com Unabridged Based on the Random House Unabridged Dictionary, © Random House, Inc. 2018\n

Prompt: We hold these truths to be self-evident, that all men (and women) are created equal, that they
Completion: are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty

Results on addition (fine-tuning)

Reconstructing n-gram non-members only take ~10 gradient steps of fine-tuning.
→ works across unseen target texts, model sizes (0.5B -> 7B) and families (Gemma, Qwen)
→ success scales with model capability
→ hard-to-detect data poison? contamination?



Different texts & configurations