Understanding the Interplay between Memorization and Learning in Large Language Models

Bishwamittra Ghosh Soumi Das Qinyuan Wu Mohammad Aflah Khan
Deepak Garg Krishna P. Gummadi Evimaria Terzi

Max Planck Institute for Software Systems | Boston University

NeurIPS 2026 | Evaluations & Datasets Track

https://arxiv.org/abs/2507.14777

"Every teacher knows that there is a profound difference between a student learning a lesson by rote and learning it with understanding, or meaningfully."

— Herbert Simon, *The Sciences of the Artificial*, 1996

Memorization and Learning in Large Language Models (LLMs)

When an LLM generates a string, can we disentangle the role of *rote memorization* and *contextual learning*?

Memorization and Learning in Large Language Models (LLMs)

When an LLM generates a string, can we disentangle the role of *rote memorization* and *contextual learning*?

- Rote memorization: Regurgitating specific details of an individual training string
- Contextual learning: Generating new strings that follow generalizable patterns of the language

When Recollection Is Not Memorization

You

`Count from 1 to 1000`

LLM

$1,\ 2,\ 3,\ \ldots,\ 1000$

When Recollection Is Not Memorization

You

Count from 1 to 1000

LLM

$1,\ 2,\ 3,\ \ldots,\ 1000$

Is this memorization or contextual learning?

When Recollection Is Not Memorization

You

`Count from 1 to 1000`

LLM

$1,\ 2,\ 3,\ \ldots,\ 1000$

Is this memorization or contextual learning?



German speaker



English speaker

Task: commit the *same* paragraph in German to memory

Who memorizes more, if recollection is equally good?

When Recollection Is Not Memorization

You

`Count from 1 to 1000`

LLM

$1,\ 2,\ 3,\ \ldots,\ 1000$

Is this memorization or contextual learning?



German speaker



English speaker

Task: commit the *same* paragraph in German to memory

Who memorizes more, if recollection is equally good?

A good measure of memorization must discount what the LLM could generate from context

Research Questions

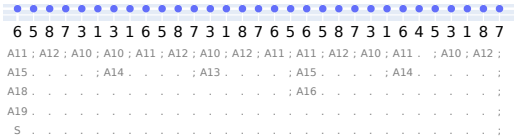
- During training, can we determine the epoch when a training string begins to be memorized?
- Can an LLM avoid memorizing any and all training strings while optimally learning a language?
- If not, can we design training schemes that reduce memorization *and* improve learning?

Research Questions & Answers

- During training, can we determine the epoch when a training string begins to be memorized?
  $\Rightarrow$ Contextual memorization disentangles rote memorization from contextual learning
- Can an LLM avoid memorizing any and all training strings while optimally learning a language?
  $\Rightarrow$ Learning a language without memorization is infeasible
- If not, can we design training schemes that reduce memorization *and* improve learning?
  $\Rightarrow$ Memorization-aware training reduces memorization and improves learning

Formal Languages as a Testbed

$S \rightarrow A19$ [1]
$A19 \rightarrow A18\ A16$ [0.95]
$A19 \rightarrow A16\ A18\ A17$ [0.05]
$A18 \rightarrow A15\ A14\ A13$ [0.95]
$A18 \rightarrow A14\ A15\ A13$ [0.05]
$A17 \rightarrow A14\ A13\ A15$ [0.95]
$A17 \rightarrow A13\ A14\ A15$ [0.05]
$A16 \rightarrow A14\ A15$ [0.95]
$A16 \rightarrow A15\ A14$ [0.05]
$A15 \rightarrow A11\ A12\ A10$ [0.95]
$A15 \rightarrow A12\ A10\ A11$ [0.05]
$A14 \rightarrow A11\ A10\ A12$ [0.95]
$A14 \rightarrow A10\ A11\ A12$ [0.05]
$A13 \rightarrow A10\ A12\ A11$ [0.95]
$A13 \rightarrow A12\ A11\ A10$ [0.05]
$A12 \rightarrow \mathbf{8\ 7}$ [0.95]
$A12 \rightarrow \mathbf{9\ 8\ 7}$ [0.05]
$A11 \rightarrow \mathbf{6\ 5}$ [0.95]
$A11 \rightarrow \mathbf{6\ 4\ 5}$ [0.05]
$A10 \rightarrow \mathbf{3\ 1}$ [0.95]
$A10 \rightarrow \mathbf{1\ 2\ 3}$ [0.05]



A probabilistic formal language L is a distribution P_L over strings

Advantages:

- Precise language specification
- Exact sampling of train and test strings
- Control over language complexity
- No semantic ambiguity, no data contamination

What do Memorization Measures Capture?

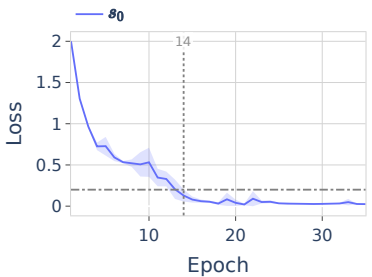
- *Privacy:* can we easily extract private information from a trained LLM?
- *Privacy:* can we infer the membership of training strings, especially rare ones?
- *Learning:* does memorization, i.e., local over-fitting, lead to suboptimal learning?

What do Memorization Measures Capture?

- *Privacy:* can we easily extract private information from a trained LLM?
  $\Rightarrow$ Recollection-based memorization
- *Privacy:* can we infer the membership of training strings, especially rare ones?
  $\Rightarrow$ Counterfactual memorization
- *Learning:* does memorization, i.e., local over-fitting, lead to suboptimal learning?
  $\Rightarrow$ No existing measure. We propose contextual memorization

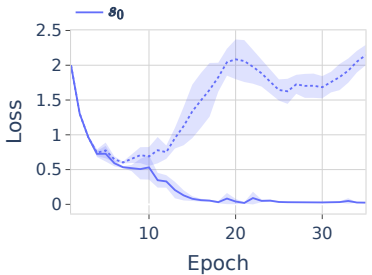
Recollection-based Memorization

Memorization $\propto$ recollection probability $= e^{-\text{loss}}$



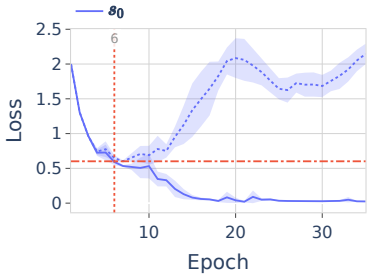
A string counts as memorized once its loss falls below a manually set threshold τ

Learning Also Leads to Recollection



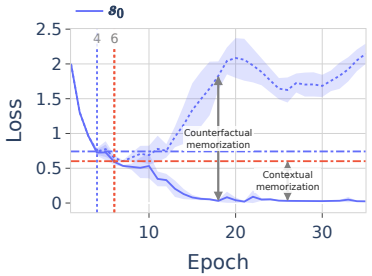
Solid line = training loss ($s \in D$) | Dotted line = contextual (test) loss ($s \notin D$)

Contextual Memorization (Ours)



Contextual memorization = the lowest contextual loss − training loss

Counterfactual vs. Contextual Memorization

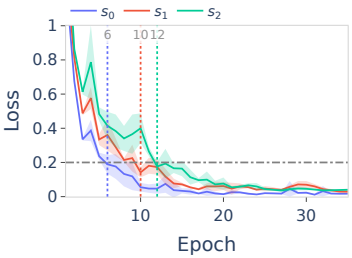


Contextual memorization is stricter than counterfactual memorization

Conceptual Comparison of Memorization Measures

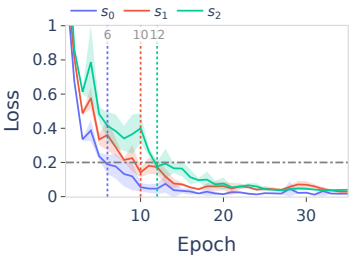
Memorization measure	Motivation	Memorization threshold	Ease of operationalization	Strictness of measure
Recollection	Disclosing private information	Manual	**Easy**	Variable
Counterfactual	Differential privacy	**Adaptive**	Hard	Medium
Contextual (ours)	Local over-fitting	**Adaptive**	Hard	**High**

Conflict of Memorization Measures

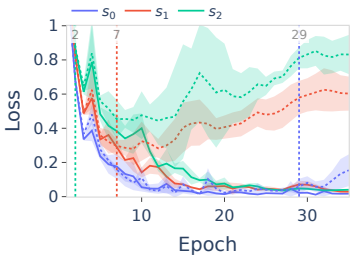


Recollection ($\tau = 0.2$)

Conflict of Memorization Measures

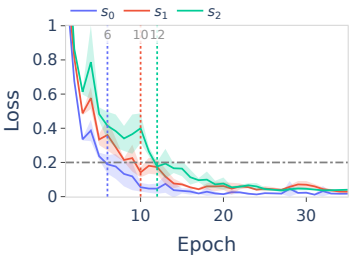


Recollection ($\tau = 0.2$)

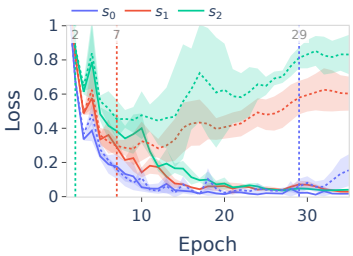


Counterfactual

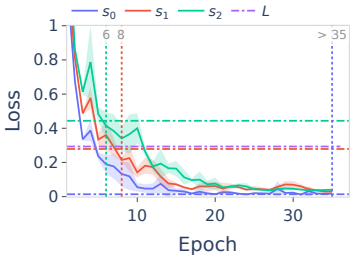
Conflict of Memorization Measures



Recollection ($\tau = 0.2$)

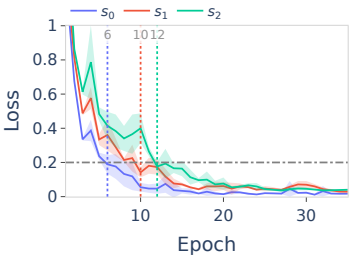


Counterfactual

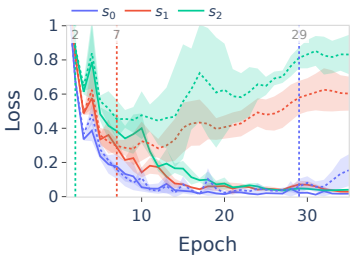


Contextual

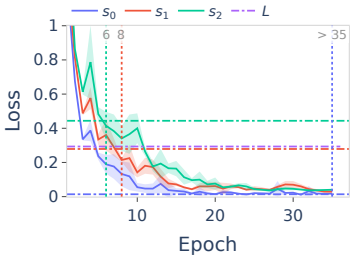
Conflict of Memorization Measures



Recollection ($\tau = 0.2$)



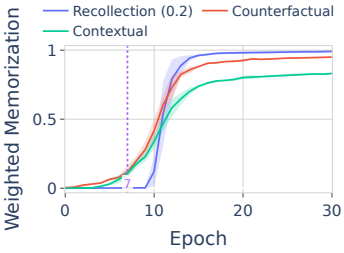
Counterfactual



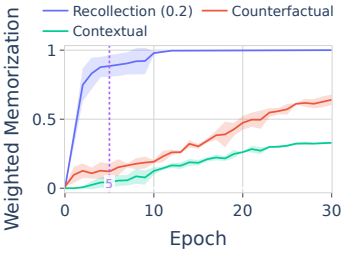
Contextual

Recollection-based memorization correlates with string frequency
Contextual and counterfactual are uncorrelated with it; here even reversed

Trade-off between Memorization and Learning

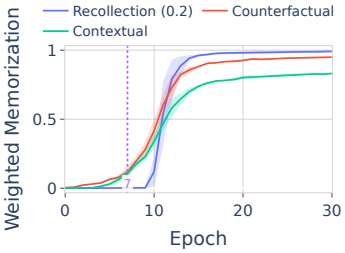


L_1 (high entropy)

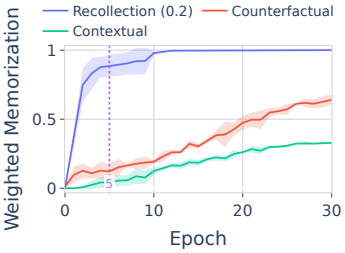


L_2 (low entropy)

Trade-off between Memorization and Learning



L_1 (high entropy)

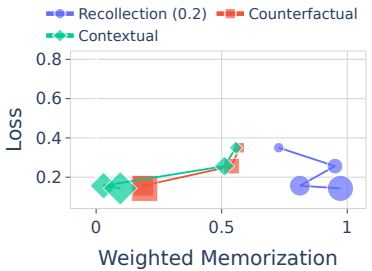


L_2 (low entropy)

Learning without memorization is infeasible in LLMs

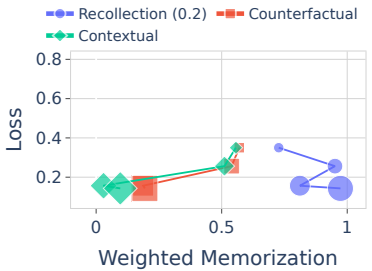
Does Better Learning Mean More Memorization?

Training size $\in \{16, 64, 256, 1024\}$ (marker size $\propto$ training size)



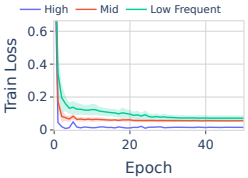
Does Better Learning Mean More Memorization?

Training size $\in \{16, 64, 256, 1024\}$ (marker size $\propto$ training size)

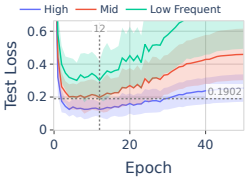


Higher training size improves learning
Better learning implies more recollection-based memorization, but less contextual memorization

Training Dynamics: Base vs. Memorization-aware Training (MAT)

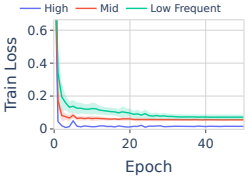


Base: train loss

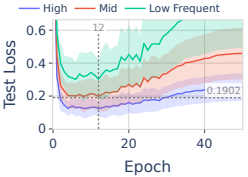


Base: test loss

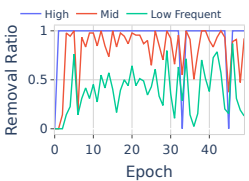
Training Dynamics: Base vs. Memorization-aware Training (MAT)



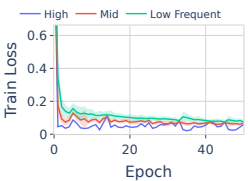
Base: train loss



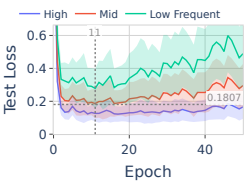
Base: test loss



MAT: strings removed

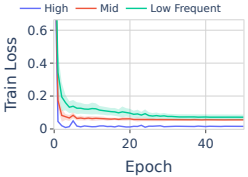


MAT: train loss

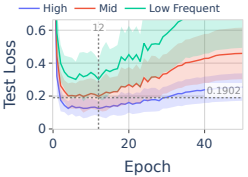


MAT: test loss

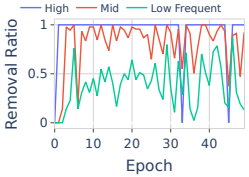
Training Dynamics: Base vs. Memorization-aware Training (MAT)



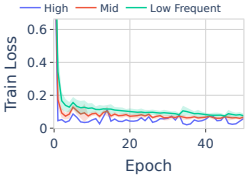
Base: train loss



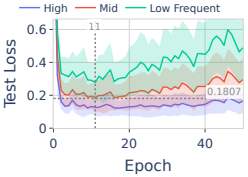
Base: test loss



MAT: strings removed



MAT: train loss



MAT: test loss

Memorization-aware training improves learning stability and uniformity

How MAT Works

At each epoch:

1. Compute $x\%$, the fraction of test strings whose loss increased since the previous epoch
2. Down-sample the $x\%$ of training strings with the lowest loss
3. A down-sampled string re-enters training if it is later forgotten, i.e., its loss rises again

How MAT Works

At each epoch:

1. Compute $x\%$, the fraction of test strings whose loss increased since the previous epoch
2. Down-sample the $x\%$ of training strings with the lowest loss
3. A down-sampled string re-enters training if it is later forgotten, i.e., its loss rises again

Analogy: memorizing the easy chapters of a book does not help you learn the hard ones.
Skip the already-learned chapters, focus on the hard ones, re-read if forgotten.

How MAT Works

At each epoch:

1. Compute $x\%$, the fraction of test strings whose loss increased since the previous epoch
2. Down-sample the $x\%$ of training strings with the lowest loss
3. A down-sampled string re-enters training if it is later forgotten, i.e., its loss rises again

Analogy: memorizing the easy chapters of a book does not help you learn the hard ones.
Skip the already-learned chapters, focus on the hard ones, re-read if forgotten.

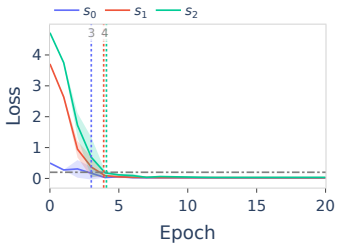
Overhead is one forward pass per epoch · fewer training strings $\Rightarrow$ lower training cost

Conclusions

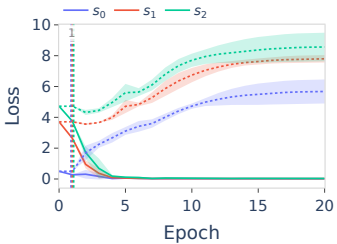
- Contextual memorization disentangles memorization from contextual learning
- Memorization is string-specific; measures disagree on when and in what order it happens
- Learning a language optimally requires memorizing some training strings
- Memorization-aware training: stable, uniform learning with lower memorization *and* lower cost

Backup slides

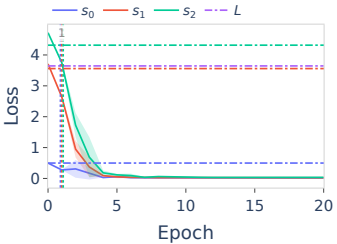
Validation: Natural Language (Tiny Shakespeare)



Recollection ($\tau = 0.2$)

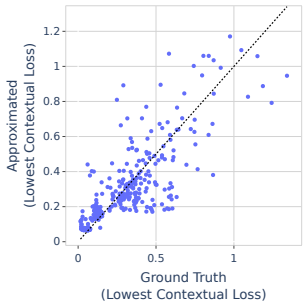


Counterfactual

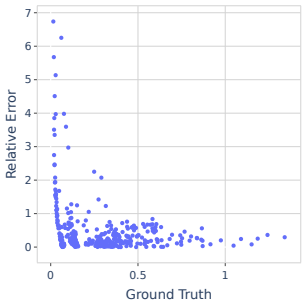


Contextual

Validation: Approximating the Optimal Contextual Loss

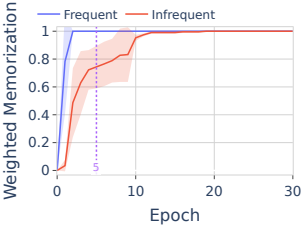


Ground truth vs. approximation

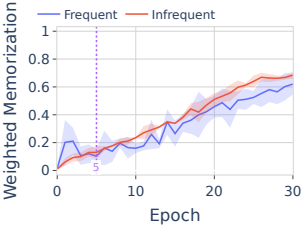


Relative error vs. ground truth

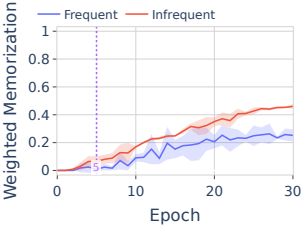
Memorization and String-Frequency



Recollection ($\tau = 0.2$)



Counterfactual



Contextual