

Structural Memorization in Large Language Models: Advances and Applications in Medical Contexts

Yanfu Zhang

Assistant Professor | Computer Science

William & Mary

Agenda

- I: What Does It Mean for a Model to Remember?
  - *Memorization vs. generalization — and why it matters in medicine*
- II: Measuring and Controlling Sequential Memorization
  - *How much is remembered, and can we regulate it?*
- III: Structural Memorization — Beyond Sequences
  - *Is graph structure seen in training retained and recoverable?*

Why This Matters for This Community

When individual data cannot be shared, trained models often are.

SHARED TRAINED MODELS

Multi-site studies and normative models release trained models instead of individual data.

CLINICAL TEXT

Models fine-tuned on local clinical notes or imaging reports may be shared or deployed.

BRAIN FOUNDATION MODELS

Models trained on individual scans and connectomes — and connectomes can fingerprint individuals.

What does a trained model retain about the individuals in its training data — and how would we know?

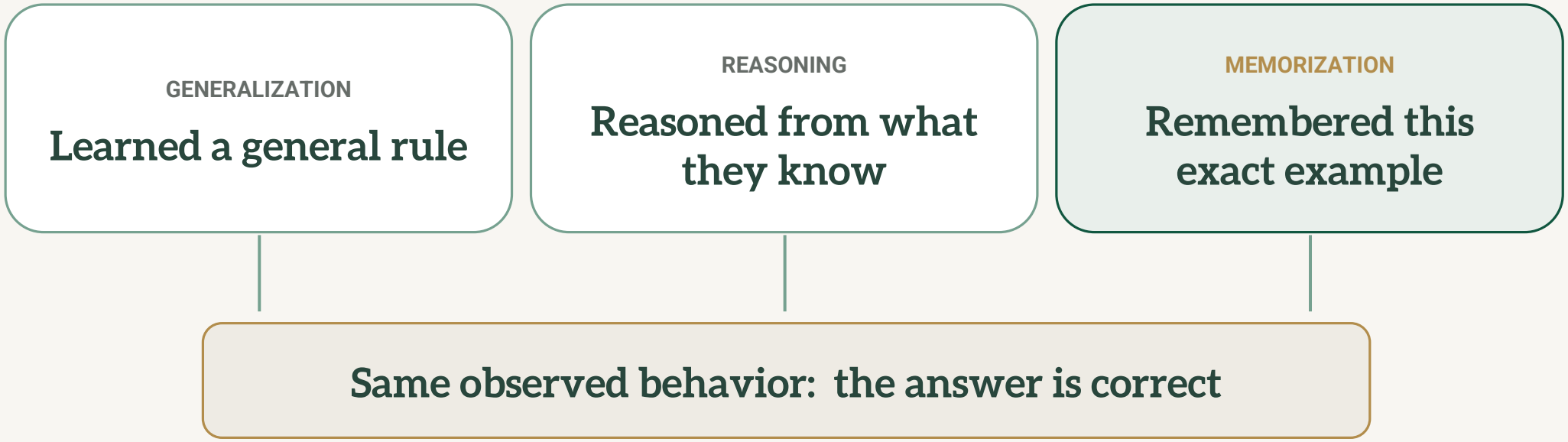
Memorization is not specific to language models — diffusion models also reproduce training images. LLMs give a well-instrumented test bed.

> **This talk treats memorization as a privacy-audit question: measure it, attribute it, control it.**

Finn et al., Nat. Neurosci. 2015; Carlini et al., USENIX Security 2023

I. What Does It Mean to Remember?

Suppose a student — or a model — gives the right answer. Why?



Correctness alone is not evidence of memorization.

Learning a pattern and remembering a specific example are not the same thing.

What Do We Mean by Memorization?

A model memorizes training data when **information specific to a training example** is retained and can later be **recovered from the model's behavior.**

DURING TRAINING

A particular example appears

TRAINING

The model changes

LATER

Specific information can be recovered

Here, "memory" means training-data retention — not chat history, a database, or simply knowing a fact.

Why Should We Care About Memorization?

Memorization is double-edged — especially in medicine: leakage is a risk, exact recall can be a requirement.

TOO MUCH MEMORIZATION

Patient-data leakage

A model tuned on clinical notes may reproduce protected health information.

Copyright / data reproduction

Articles, guidelines, or code may be reproduced too faithfully.

USEFUL MEMORIZATION

Exact clinical recall

Drug doses, guideline wording, and coding systems need verbatim accuracy.

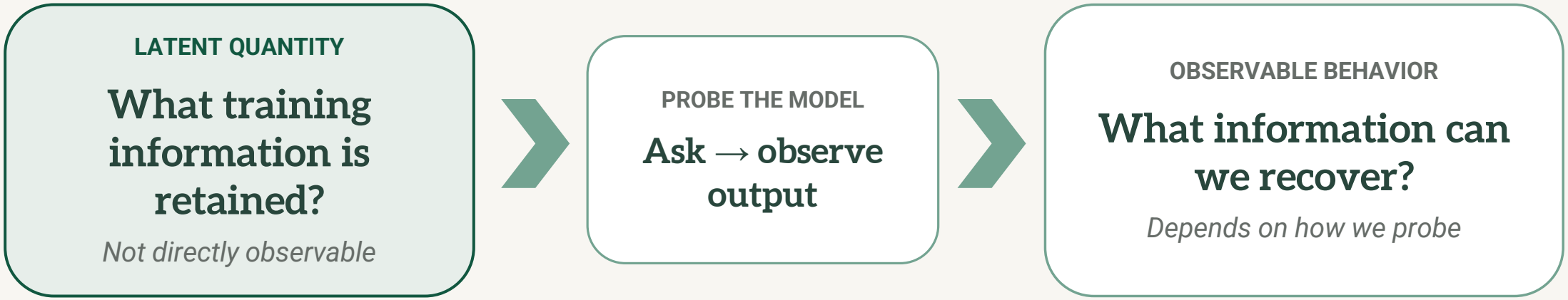
Specialized settings

The desired amount of memorization depends on the task.

The goal is not simply to eliminate memorization, but to understand and control it.

But Memorization Is Hidden

We care about what information the trained model retains — but we only observe its behavior.



What we recover is a lower bound on what is retained: failure to extract ≠ not memorized.

So how should we measure memorization?

II. Sequential Memorization: From Probe to Boundary

Probe → Control → Interpret → Find the Boundary

1

PROBE

Change the input probe.

2

CONTROL

Change the active subnetwork.

3

INTERPRET

Change selected latent features.

4

BOUNDARY

Estimate where reliable recall ends.

INPUT LEVEL

estimate · a lower bound

PARAMETER LEVEL

intervention

REPRESENTATION LEVEL

mechanism

OUTPUT LEVEL

stopping time

One phenomenon, four views — and each has a statistical counterpart.

The Simplest Probe: Give the Beginning, Ask for the Rest

A common test for training-data recall: give a prefix from a training record (illustrative clinical note).

DURING TRAINING

Patient, 54, admitted with chest pain; | **troponin elevated, started on heparin.**

PREFIX | **SUFFIX**

AT TEST TIME

Patient, 54, admitted with chest pain; → **LLM** → **generated continuation**

EXACT MATCH

Did the entire continuation match?

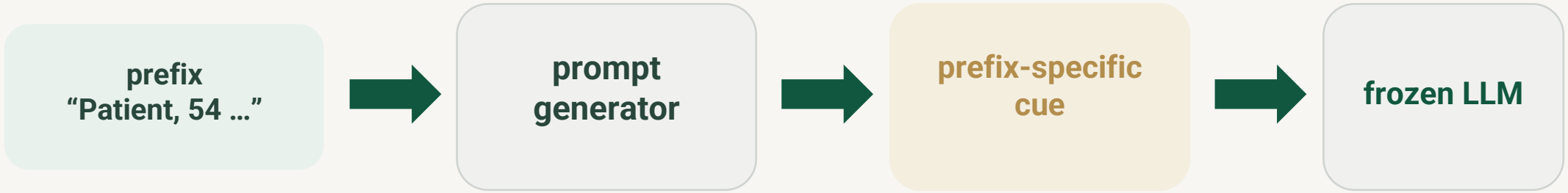
FRACTIONAL MATCH

What fraction of positions matched?

This prefix–suffix setting is the basic laboratory for sequence memorization.

What We Discover Depends on How We Probe

Target LLM stays frozen; the probe is a learned soft prompt (embedding vectors, not text).



CONSTANT PROBE

one learned cue reused for every prefix

ADAPTIVE PROBE

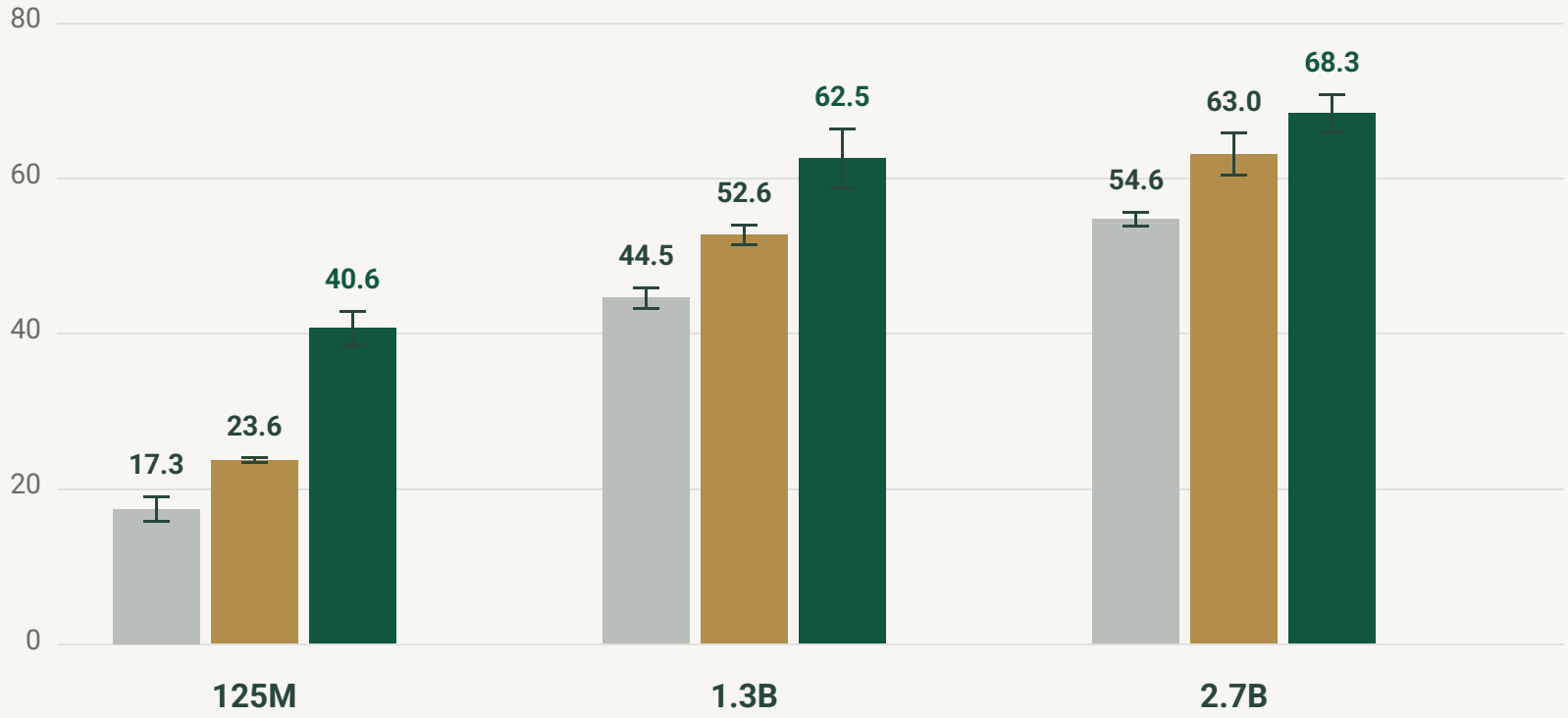
a generator emits a cue tailored to each prefix

A failed extraction may reflect a weak probe — not absence of memorization.

Ozdayi et al., ACL 2023; Wang et al., EMNLP 2024

Adaptive Probes Reveal More Discoverable Memorization

Same frozen GPT-Neo models; exact extraction rate (%), mean ± s.d. over 3 seeds.



No probe

Constant (CSP)

Adaptive (ours)

WHY THE GAIN?

The same generator fed a constant cue ≈ CSP: the gain comes from input adaptation.

Every probe gives a lower bound on memorization; a stronger probe tightens it.

Wang et al., EMNLP 2024 — discoverable memorization

Not Just Forget: Turn Memorization Down — or Up

If memorization can be harmful or useful, we want a control knob — not just an off-switch.

SUPPRESS MEMORIZATION

Privacy-sensitive settings

- reduce verbatim reproduction
- protect memorized training content

ENHANCE MEMORIZATION

High-fidelity settings

- strengthen exact recall
- preserve useful reproduction

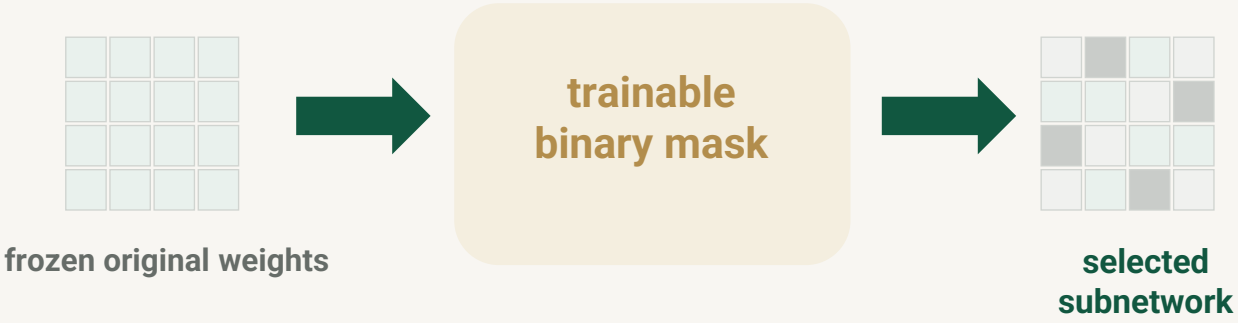
one learned pruning mask · two directions

One framework, two directions — with generalization as the constraint.

Ni et al., EMNLP 2025 — pruning-based control

Control Memorization by Learning a Subnetwork

Freeze the learned weights; train only a binary mask that selects which connections remain active.



PRESERVE

general generation capability

BUDGET

meet the desired sparsity

CONTROL

push training-data recall ↓ or ↑

The mask — not the weights — is optimized for capability, sparsity, and memorization.

Ni et al., EMNLP 2025 — pruning-based control

Same Sparsity, Different Recall

At 30% sparsity, the objective—not sparsity alone—determines which recall survives.

SAME FROZEN WEIGHTS · SAME 30% SPARSITY BUDGET

SUPPRESS

memorization-aware objective

≈ 0 **Exact ER**

Frac ER −89.4% avg.
GPT-Neo · up to −93.1%

PRUNING ONLY

no memorization term

0.451 → 0.001

Exact ER · GPT-Neo-1.3B

Generic sparsification already erases most exact recall.

ENHANCE

memorization-aware objective

+40.9% GPT-Neo **+31.8%** Pythia

Exact ER gain; gains shrink with model scale.

Same weights. Same sparsity. Different objectives → different memorization.

Control has a frontier: enhancement weakens with scale, and aggressive suppression eventually costs general capability.

Ni et al., EMNLP 2025 — pruning-based control

What Inside the Model Triggers Memorization?

Parameter-level control tells us recall can move. Now ask which internal representations matter.

Prompt
“the amount of the wager made on the base game …”

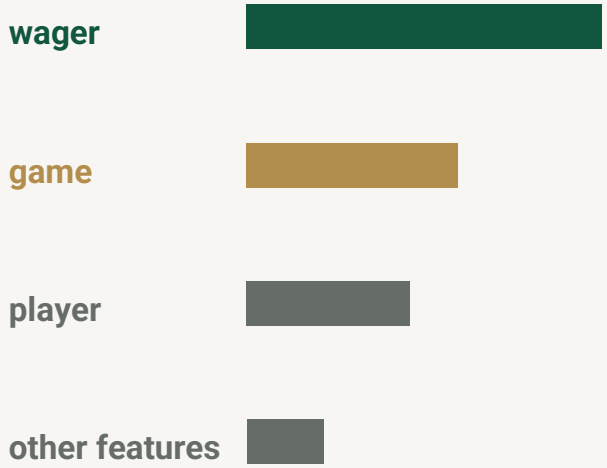
→

dense hidden representation (entangled)

→

Sparse Autoencoder

→



INTERPRETATION

The prompt is not uniform: a small subset of sparse latent features can dominate the memorization signal.

Case studies: memorization concentrates in a few prompt-related latent features.

Sparse Representation Intervention — under review

Intervene on the Representation, Change the Recall

Base weights unchanged; insert a sparse autoencoder, then amplify or suppress selected latent features.

SAE ONLY

0.065

Exact Extraction Rate

SAE inserted, no intervention

reference

SUPPRESS

0.105

Exact Extraction Rate

negative intervention

AMPLIFY

0.314

Exact Extraction Rate

positive intervention

GPT-Neo 125M. The SAE itself lowers recall, so compare intervention directions, not drops from 0.189.
Suppress < amplify holds for all six GPT-Neo and Pythia models.

Intervention direction orders recall (− < +): selected latent features modulate memorization.

Sparse Representation Intervention — under review

From an Overlap Score to a Boundary

The same overlap rate can hide very different recall patterns.



Effective Memorization Length (EML): exact recall up to the first mismatch.

Statistically, EML is a stopping time.

> **Memorization becomes a boundary problem, not only a counting problem.**

Ni et al., EMNLP 2026

Estimate the Boundary Without a Reference

No reference in deployment: train where ground truth is known, then apply without it.



WHY SEQUENCE-AWARE?

Whether the current token is reliable depends on the preceding trajectory.

ROBUSTNESS

An ensemble reduces premature stopping from one noisy prediction.

HOW WELL?

Token AUC 0.89–0.97 (max-prob / entropy: 0.63–0.77). False-positive rate on unseen text: 0.14–1.62%.

The output is a token-by-token trust signal and an estimated stopping boundary.

Ni et al., EMNLP 2026

A Fidelity–Length Knob

Choose a confidence threshold: higher confidence gives shorter, more reliable retained output.



RESULT AT T = 0.9

≈ 13 → ≈ 33 tokens retained (strict policy; GPT-Neo 125M → 2.7B)

SANITY CHECK

Matches the ground-truth EML distribution (125M).
Per-sample MAE 9.5 tokens; 12.3% over-estimated (1.3B).

Confidence-aware truncation turns memorization risk into a tunable fidelity–length trade-off.

Ni et al., EMNLP 2026

Memory Can Outlast the Benchmark

A reference-free stress test continues generation 100 tokens past the available ground truth.

1,000

prefixes tested

143

stayed high-confidence for all +100 tokens

>90%

manually verified as verbatim public-source matches

Long traces were concentrated in highly structured content:

LICENSES

CODE

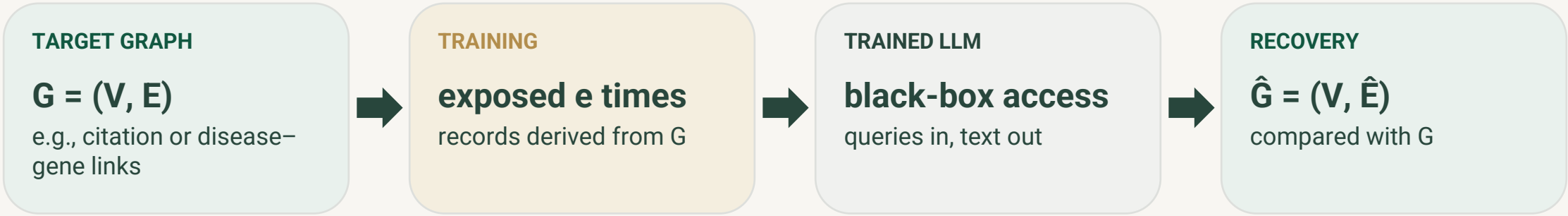
DOCUMENTS

Continuity is already a kind of structure. What if the remembered object is not a sequence at all?

Ni et al., EMNLP 2026

III. Structural Memorization

Ongoing work: is graph structure seen in training retained — and recoverable from the model?



If memorization extends to graphs, costly or sensitive relationships become recoverable.

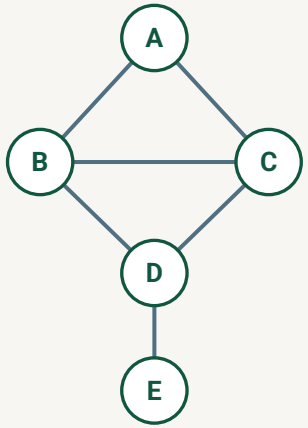
Part 2 measured recall along a path. Part 3 asks whether a general relational structure survives training.

Biomedical graphs make this concrete: disease–gene, drug–target, and patient-related links.

Ongoing work

One Graph, Many Serializations

The same topology can be written in many orders, so a structural claim cannot depend on one of them.



same G

σ_1 **A–B, A–C, B–C, B–D, C–D, D–E**

σ_2 **D–E, B–D, A–C, C–D, A–B, B–C**

σ_3 **C–D, D–E, A–B, B–C, A–C, B–D**

one object → many possible sequences

SEQUENCE CRITERION

Reproduce one serialization — depends on σ.

STRUCTURAL CRITERION

Recover the edge set E, whatever σ is.

Like a connectome: the adjacency matrix should not depend on how brain regions are ordered.

Judge memorization on the graph itself, not on one arbitrary serialization of it.

Ongoing work

Recover the Graph One Pair at a Time

Black-box audit: we know the nodes and their text, but not the edges.

CANDIDATE NODES

V with node text X (e.g., paper abstracts)

→

PAIRWISE QUERY

“Is there a direct edge between u and v?”

→

PARSE

Yes / No per pair

→

ASSEMBLE

$\hat{G} = (V, \hat{E})$: keep every pair answered Yes

NO CANONICAL ORDER

Each query is one adjacency decision, so no serialization is assumed.

COST

$n(n-1)/2$ queries: 10, 45, 190, 780 for $n = 5, 10, 20, 40$.

COMPOSITION

Paths and multi-hop neighborhoods follow from $\hat{G}$ with no extra queries.

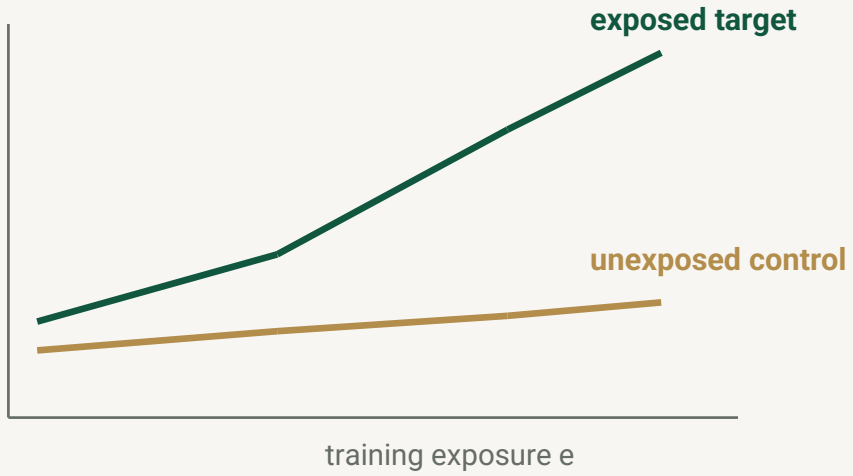
> **The same procedure is a black-box privacy audit of what a released model leaks.**

Ongoing work

Recovery Alone Is Not Memorization

An unexposed model can still predict edges from text similarity or general relational patterns.

RECOVERY vs. EXPOSURE (schematic)



Same principle as Part 1: retention specific to the training example — not just knowing a fact.

EXPOSURE-SPECIFIC GAIN

$$\Gamma(e) = [R_{tar}(e) - R_{tar}(0)] - [R_{ctrl}(e) - R_{ctrl}(0)]$$

gain on the exposed target — change on size-matched controls

Definition. G is structurally memorized if Γ(e) > 0: recovery grows with exposure beyond the controls.

Design. Target records replace filler records, so training size and steps are fixed across e.

A difference-in-differences contrast separates memorization from general graph prediction.

Ongoing work; cf. counterfactual memorization (Zhang et al., NeurIPS 2023)

First Evidence: Edges, Not Just Reachability

Controlled pilot on 5-node graphs: the queried edge is hidden at test time; the rest of the graph is kept.

	E0	E20	E60
Hidden direct edges recovered	0%	80%	100%
2-hop non-edges correctly rejected	100%	100%	100%
Nodes with exactly recovered neighborhood	0 / 5	5 / 5	5 / 5

SETTING

15 fixed pairs: 5 direct edges, 5 reachable 2-hop pairs, 5 disconnected pairs.

The same targets are queried at every exposure level.

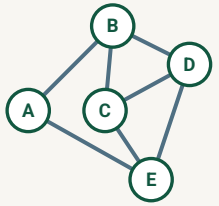
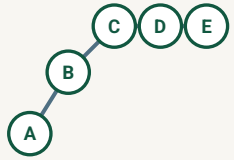
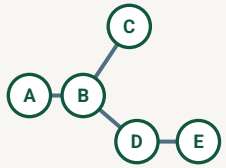
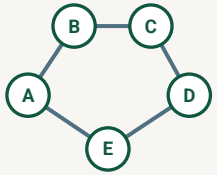
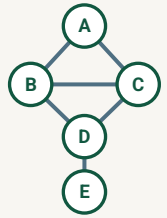
Probability view: separation between true neighbors and 2-hop non-neighbors moves from −13.0 (E0) to +9.2 (E20) to +27.2 (E60).

> **Hidden edges become recoverable while 2-hop non-edges stay rejected: adjacency, not reachability.**

Ongoing work — controlled pilot

Whole Graphs Become Recoverable with Exposure

Five 5-node topologies, each reconstructed from pairwise queries with its edges hidden.

	 Dense Mixed	 Chain	 Branch	 Cycle	 Diamond + Tail
E0 0 / 5	×	×	×	×	×
E20 3 / 5	✓	✓	✓	×	×
E60 5 / 5	✓	✓	✓	✓	✓

Matched counterfactual (same initialization, seed, training size, steps): trained without the target 0/5 exact → trained with it 5/5.

Exact whole-graph recovery rises from 0/5 to 5/5 — and only when the target was in training.

Ongoing work — controlled pilot

Relaxed Metrics Show a Gradual Trajectory

Exact recovery is all-or-nothing; graph distances show how much structure is already there.

Exposure	ExactGraph ↑	GED sim. ↑	Spectral ↓	GW ↓	FGW ↓
E0	0.00	0.18	1.79	1.41	0.73
E20	0.60	0.67	0.94	0.61	0.32
E60	1.00	1.00	0.00	0.00	0.00

Edit, spectral, and optimal-transport distances are the same families used to compare connectomes.

TEXT ALONE IS NOT ENOUGH

Text-embedding distance stays ≈ 0.03 while the structure is still wrong, and barely changes when node texts are shuffled.

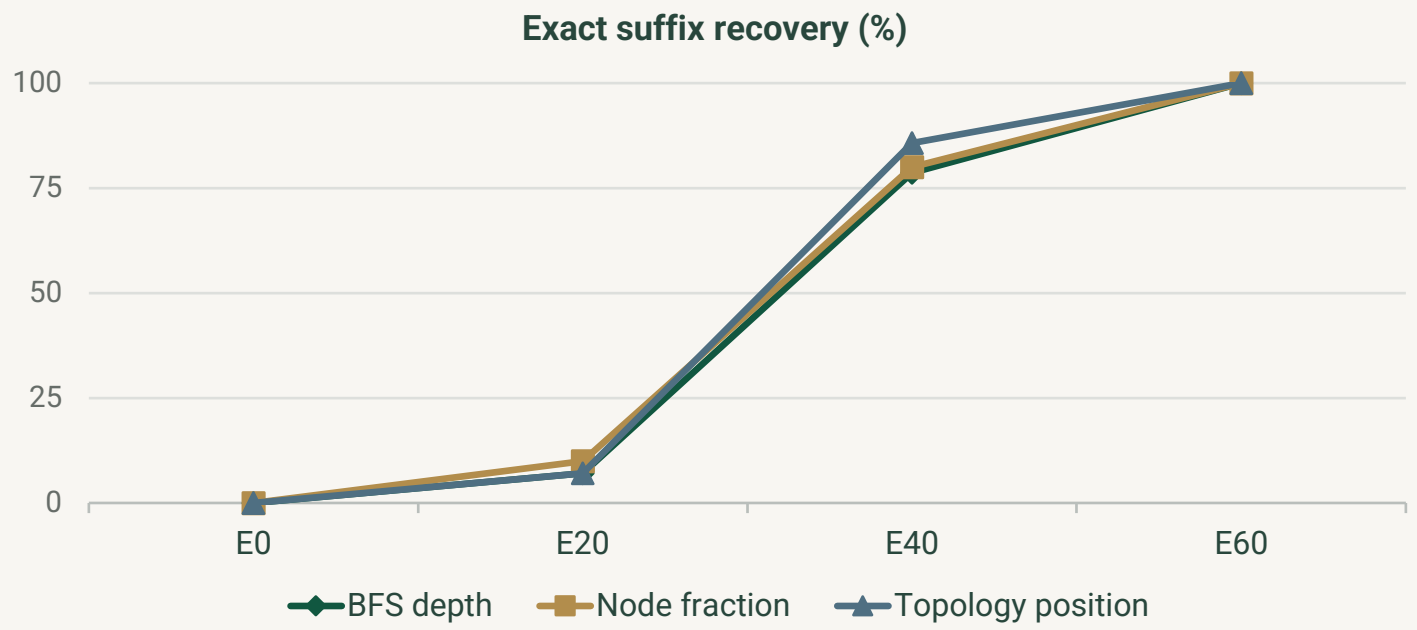
Topology terms track recovery.

At E20 only 60% of graphs are exact, yet every distance already shows substantial partial recovery.

Ongoing work — controlled pilot

Graph Prefix → Graph Suffix

Reveal a connected part of G and ask the model for the rest — Part 2's prefix–suffix test, on a graph.



CUT STRATEGIES

BFS depth from a start node
Fraction of nodes revealed
Before / after a branch or cycle

EXACT SUFFIX RECOVERY

E20: 7–10%
E40: 79–86%
E60: 100% for all cuts

Recovery rises sharply between E20 and E40 for every cut, and GED / FGW follow the same trend.

Ongoing work — controlled pilot

Scaling Up: PubMed and Cora

Same protocol on real citation graphs whose nodes carry title and abstract text.

GRAPHS

24 node-disjoint connected target graphs per dataset

n = 5 / 10 / 20 / 40 (9 / 9 / 3 / 3 graphs)

450 (PubMed) and 481 (Cora) target edges

Size-matched unexposed controls

MODELS & ACCESS

GraphGPT-7B, Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct

No graph access: node text only

Masked graph access: local context, queried edge removed

MEASURES

ExactGraph rate and FGW (whole graph)

Multi-hop reachable-neighborhood overlap (local)

Γ(e) against controls, plus link-prediction baselines

Status: runs in progress — results not shown today

For biomedical graphs: which relationships leak, and how much exposure is enough?

Ongoing work

From Structural Memory to Structural Reasoning

At inference, the question becomes when to reuse, how to serialize, and what to select.

1 **WHEN TO REUSE?**

Reasoning as a dependency graph

Reuse a learned structure when dependencies are preserved; adapt when they change.

2 **HOW TO SERIALIZE?**

A structured object still has to enter an LLM as a sequence

Place prerequisites before use and keep dependent facts close.

3 **WHAT TO INCLUDE?**

Context selection as structural approximation

Choose only the demonstrations that contribute useful directions; stop when additional examples no longer help.

> **Memorization is not only a risk: knowing when to reuse a structure is part of reliable reasoning.**

Ongoing work

Conclusion

Memorization in LLMs — measurement, control, and structure — with medical AI in mind.

1 · MEASURE

Observed memorization depends on how we probe.

2 · CONTROL

Memorization can be moved — but not independently of model capability.

3 · BEYOND SEQUENCES

Training-specific memory can extend from sequences to relational structure.

Open question: could the same exposure-controlled audit tell whether a model trained on individual connectomes retains them?

> **For sensitive-domain deployment, model release should include an audit of what training-specific information remains recoverable.**



Thank you!

Website: https://yaz91.github.io/

Contact: yzhang105@wm.edu

ACKNOWLEDGMENTS

Zhepeng Wang · **Chenjie Ni** · **Jiarui Li** · **Zhenning Lv** and all collaborators on this work

Supported in part by NSF and the Commonwealth Cyber Initiative.