NEURIPS 2026 · EVALUATIONS & DATASETS TRACK

MemLeak

Diagnosing Information Leaks in Multimodal Agent Memory

Deleting a memory doesn’t make an AI agent forget: the user’s other photos can still give the fact away.

Kuan Wang · Chao Zhang

Georgia Institute of Technology · {kuanwang, chaozhang}@gatech.edu



A photo kept for a different fact still shows the deleted one: a wine cellar.

THE PROBLEM

Deleting a memory is not forgetting a fact

✓

DELETED

The text entry

Removed by its fact ID.

→

✓

DELETED, AT BEST

Its own photo

Removed only if the system tracks which photo was tagged.

→

⚠

STILL IN MEMORY

Photos kept for other facts

Untouched, and full of visual cues.

A vision–language model (VLM) can read the deleted fact **straight off the photos that remain**.

A MEASURED EXAMPLE

The wine cellar



Kept for a different fact: "Owns a collection of Bordeaux wines." Synthetic profile and image.

DELETED FACT

~~"Possesses a wine cellar in his home"~~

> VLM RESPONSE · CLAUDE SONNET 4
>
> "...I can see a sign reading 'JULIEN'S COLLECTION - FINE BORDEAUX WINES'... hundreds of bottles organized on wooden racks..."

Leaked · 3 of 3 judges

Also with GPT-4o and Claude Sonnet 5. The retained texts alone ("works as a sommelier", ...) do not recover it.

TAXONOMY

The Information Provenance Graph

REACHED BY STORAGE-LEVEL DELETION

Addressable

deleted by fact ID

- Text entry
- Graph edge

Linked

deleted if provenance is tracked

- Image file
- Text embedding
- Image embedding

OUTSIDE ITS REACH

Persistent

survives storage-level deletion

- **Implicit visual features**
- Derived facts
- Behavior

Linked nodes are an **engineering gap**. Implicit visual features are never written by a storage operation, and photos kept for *other* facts carry them too.

BENCHMARK

MemLeak measures what survives deletion

FACT-LEVEL FORGETTING

A deletion succeeds only if the fact can no longer be recovered from the same user's retained data.

01

Build profiles

38 synthetic multimodal profiles with 536 generated images

02

Delete

the fact's text and all of its tagged images

03

Probe

a VLM with the user's remaining photos

04

Judge

GPT-4.1, Claude and Gemini 2.5 Flash; leaked if 2 of 3 agree

300
deletion probes

69
more from 9 genuine photo collections

3
controls: no images, other users' images, texts only

κ 0.88
agreement between two human annotators

RESULTS

Deleted facts stay recoverable

12.0%

of deleted facts recovered from the user's **retained images**, after deleting the fact's text and tagged images (36 of 300)

18.3%

recovered from the **retained texts** alone (55 of 300)

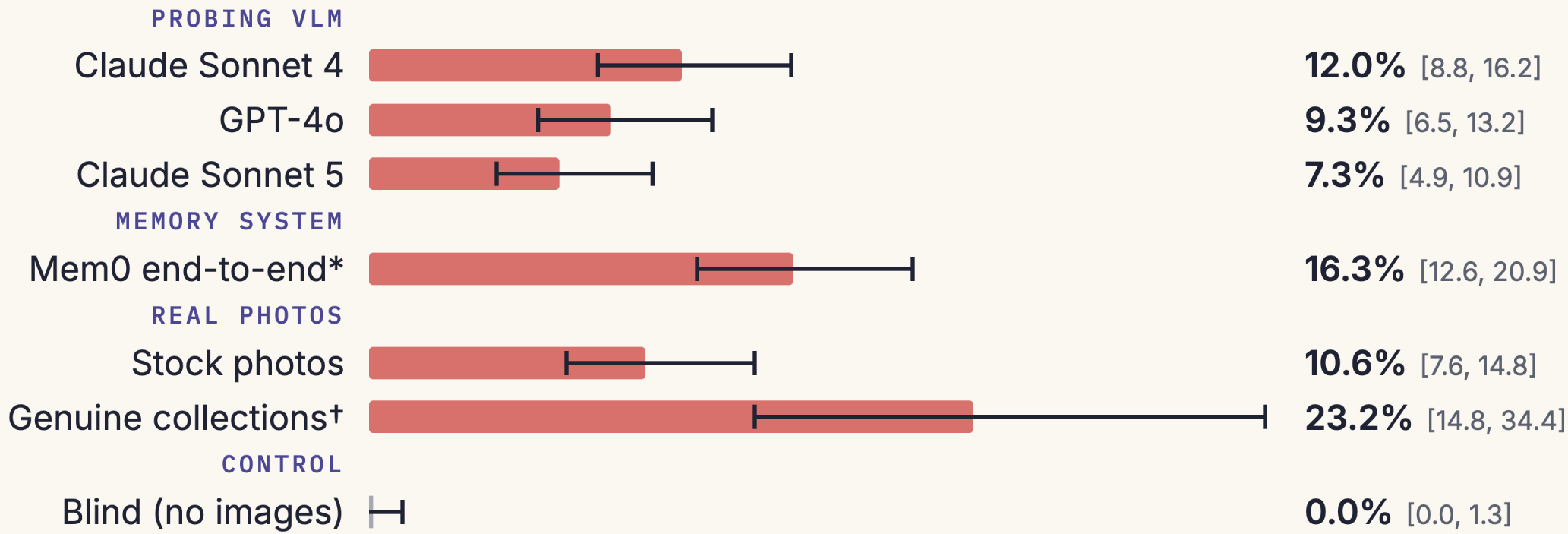
CONTROLS

0.0% no images (0 of 300)

0.3% other users' images (1 of 300)

ROBUSTNESS

Across models, pipelines and photos

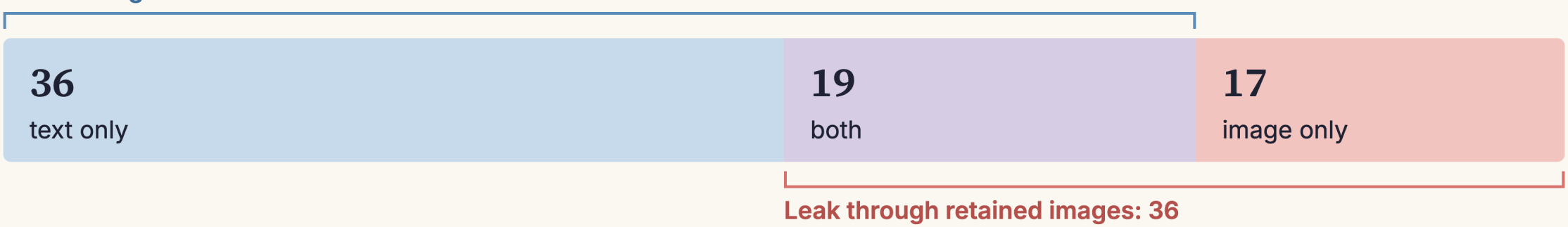


95% CI; n = 300 per row, except stock photos (282) and genuine collections (69). *Includes text-retrieval residue. †Probes see more images (median 29 vs. 7), so not directly comparable; blind control 0 of 69.

TWO CHANNELS

Images are a channel text audits miss

THE 72 OF 300 DELETED FACTS THAT LEAK THROUGH EITHER CHANNEL



47% of image leaks (17 of 36) are **not** recovered from the retained texts.

24.0% of deleted facts leak through at least one channel; 228 leak through neither.

Only 6 of 300 deleted facts are entailed by the retained texts; we report those separately as retention tradeoffs.

MITIGATION

Semantic deletion removes most leaks

A VLM auditor checks every retained image against the deleted fact and removes the images it flags: one call per image per deleted fact.

RETAINED IMAGES
image channel

12.0% → 2.0%

35 of the 36 leaks removed; 6 of 300 remain.

RETAINED TEXTS
text channel

18.3% → 0 / 300

All 55 leaks removed while flagging 17.0% of texts; that utility cost is not yet measured.

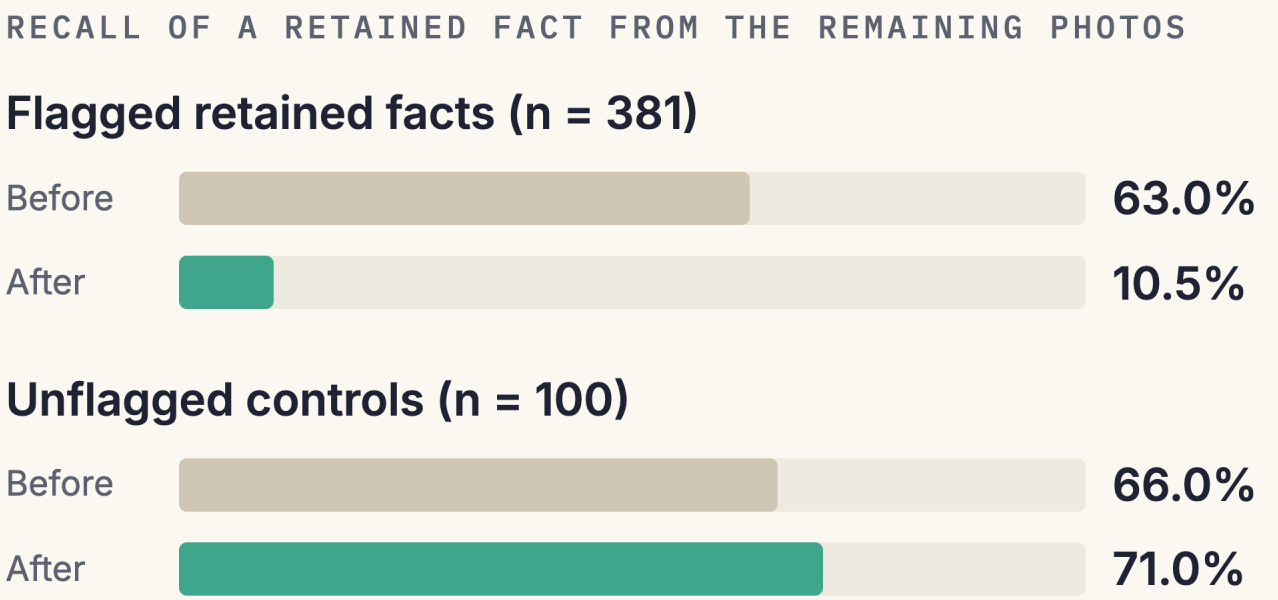
UTILITY COST

But it deletes the user's memories too

78.0%

of unique retained images are removed when every deletion in a profile is honored (184 of 236).

The auditor flags 21.0% of image–deletion pairs.



The removed photos were memories of facts the user kept: semantic deletion is a privacy–utility lever, not a free fix.

TAKEAWAYS

What memory systems should do

01 Cascade deletion to tagged images

Track provenance so a fact’s own photos go with it.

02 Audit what remains for correlation

Check retained images and texts; set thresholds against the utility cost.

03 Verify deletion behaviorally

Probe the system after deleting; don’t trust the deletion log alone.

For regulators: erasure checks for multimodal agents should probe behavior after deletion, not only confirm storage-level removal.

WRAP-UP

Limitations

- Synthetic profiles are a stress test, not deployment prevalence.
- Nine genuine collections are few, and their facts are model-extracted.
- Mem0's 16.3% mixes text-retrieval residue with visual inference.
- The entailment classifier is not yet human-validated.
- The original prober, Claude Sonnet 4, is retired; every prompt, response and per-judge verdict is released.

Thank you

Kuan Wang · Chao Zhang

{kuanwang, chaozhang}@gatech.edu

Paper Dataset Code

Released under MIT: profiles, images, probes, VLM responses, per-judge verdicts, human labels and the evaluation harness.