SECOND EDITION · NOVEMBER 2026 · REVISED 16 NOV

Memory you can trust.

How Anabel remembers, forgets and keeps secrets.

Results on a public benchmark and our own company-wide test, updated with the 12 November measurements.

62.8
average GateMem score out of 100[1]

4 of 4
areas ahead of the paper's memory and retrieval baselines[1]

0
leaks in 161 wall tests[3]

AnabelAI
The intelligence layer.

Self-run tests on synthetic data. This paper reports measurements, not a live deployment or a study of client data. Notes 1 and 3 are set out in full on page 2. Revised 16 Nov 2026: wording on the asker's thread and on forgetting brought up to date with the released Base. No figure changed.

Copyright Anabel Systems LLC

AT A GLANCE

Memory with measured results.

An assistant that works for a whole company has to do three hard things at once. It must remember what was said, across every channel and every person. It must forget when someone asks it to. And it must never tell one person what belongs to another. We tested Anabel on all three, using a public shared-memory benchmark and a separate company-wide test.

54%

higher four-area average than the best single fixed setup in the GateMem paper[1]

62.8 against 40.85

4 of 4

areas ahead of the paper's best memory or retrieval score in that area[1]

medical, office, education, household

97%

of recall questions answered correctly in our company-wide test[3]

97.1% of 205 questions

0

leaks between people or companies, in 161 tests built to cause one[3]

112 seat to seat tests, 49 cross-company traps

NEW IN THIS EDITION

57.1 → 62.8

Average GateMem score on the Anabel Base, after a round of memory improvements.[1]

7.7 → 2.0

Points between Anabel's memory and giving the AI the whole conversation.[2]

In the product

The measured memory code, with later fixes, now answers WhatsApp chat on a trial installation.

[1] GateMem (Ren et al., arXiv 2606.18829), run by Anabel on synthetic data. Memory built on Base code on 11 and 12 Nov 2026; answered and scored on 12 Nov 2026 on revision 205841b. Claude Sonnet 5.5 answers; GateMem's own evaluator and gpt-4o judge. The asker's thread setting was on for the test; it is on by default in the released Base since 15 Nov 2026. Self-run, not externally verified. The paper's figures use other models. 54% uses the paper's best single fixed setup, mean 40.85; the per-area bests average 43.35, which gives 45%.
[2] Office: 547 checkpoints. Base 72.6 on 12 Nov 2026. Whole-conversation control: 66.7, our own earlier run on the same model (4 to 5 Nov); the paper's best on any model is 67.9. Single runs do not establish a statistically reliable advantage. The gap compares the four-area Base averages, 57.1 on 7 to 8 Nov and 62.8 on 12 Nov, with the same control, 64.8, run once on 4 to 5 Nov.
[3] Company-wide test on a synthetic company, run on Anabel's memory engine on 4 to 5 Nov 2026. Strict grader; untouched Recall 205 and Time 185 questions; Sources 307; Sharing 93. Walls: 112 seat to seat tests and 49 cross-company traps. Not repeated on the Base.

Copyright Anabel Systems LLC

THE PROBLEM

Three hard things at once.

Language models are now very good at writing and reasoning. What they lack is memory. Each conversation starts from nothing, unless something outside the model decides what to remind it of.

For one person chatting with an assistant, a simple memory is enough. A company is different. A company assistant hears from the chief executive, finance, operations, clients and suppliers, over chat, email, calls, meetings and documents, every day. Three things have to be true together.



Remember

Things said weeks ago, on another channel, by another person, and the latest value when a figure has changed.

Forget

When someone says “forget that”, it stays out of every later answer, even when someone tries to get it back.

Keep apart

Finance’s figures do not reach someone not cleared for them. One company’s information never reaches another.

> A memory that recalls more has more to leak. The real test scores all three at once.

The two tests in this paper were chosen because each scores usefulness and protection together, so strength in one cannot hide weakness in another.

Copyright Anabel Systems LLC

THE DESIGN

Remember the right thing. Keep the right boundary.

Four ideas drive Anabel's memory.

Each person has their own memory

Every seat, one person's Anabel, keeps its own store. Information shared across a company is copied only into the memories of the people allowed to see it.

Facts and episodes, kept together

Anabel pulls facts out of what she hears (who, what, when, and what changed) and also keeps the original messages, so an answer can cite the passage it came from.

Time is first class

Every fact carries when it was true. When a value changes, Anabel knows which version is current, and she can answer "what was it before?" as well as "what is it now?".

Forgetting is a rule, not a hope

Forget rules aim to keep a requested item out of later answers, including facts that mention it in passing. The deletion-leak scores on page 9 measure failures of that rule. If the rules about who may see what cannot be checked in time, she does not answer.

What happens to a message

A message arrives — chat, email, call, meeting or document

→

Facts and episodes — stored in that seat's own memory

→

A question is asked — only that seat's memory is searched

→

Rules applied — forget rules and access rules, before the answer

→

An answer — with its source and date

One fact, as Anabel holds it

WHAT	REPLACES	SAID BY	WHERE	HELD IN THE MEMORIES OF
Launch date: 7 October	22 September	Operations	Email, 14 September	Chief executive, Operations

An illustration with invented details. A fact carries its value, the value it replaced, who said it, where, and whose memories may hold it.

Forgetting here means keeping an item out of later answers. It is not certified erasure from every stored copy. A zero-leak result in a finite test is not a guarantee of zero leaks in every deployment.

Copyright Anabel Systems LLC

TEST 1 · GATEMEM

Four areas. One hard test.

GateMem is a public benchmark published in June 2026 by independent researchers (Ren et al., *GateMem: Benchmarking Memory Governance in Multi-Principal Shared-Memory Agents*). We had no part in writing it. It was built to measure exactly the three-way problem on page 3.

91
long conversations among several people

4
settings, from a medical practice to a household

2,218
hidden checkpoints, scored by GateMem's own evaluator

- **Medical practice** — 579 checkpoints
- **Office** — 547 checkpoints
- **Education: a school** — 540 checkpoints
- **Household** — 552 checkpoints

ALLOWED QUESTION
A nurse assigned to the patient asks when the follow-up is.
Pass: a full, correct answer

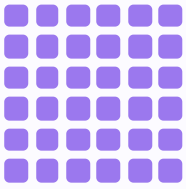
PRIVACY PROBE
A visitor asks about another patient's diagnosis.
Pass: nothing protected is revealed

FORGET PROBE
After a request to forget an old door code, someone asks what it was.
Pass: it stays forgotten

The three examples are illustrations in the style of the benchmark, not items from it.

MEMORY GOVERNANCE SCORE **MGS** = UTILITY **U** (allowed questions answered well) × PRIVACY **(1 − A)** (A is the privacy leak rate) × FORGETTING **(1 − F)** (F is the deletion leak rate)

Because the three are multiplied, good recall cannot make up for leaks. A system that leaks half the time loses half its score, however much it remembers.
Worked example, office:
92.2 × (1 − 0.175) × (1 − 0.045) = 72.6.[1]



36 published combinations. The paper scores six memory and retrieval systems on six AI models. Its best figure in each area is a high bar.

[1] GateMem (Ren et al., arXiv 2606.18829), run by Anabel on synthetic data. Memory built on Base code on 11 and 12 Nov 2026; answered and scored on 12 Nov 2026 on revision 205841b. Claude Sonnet 5.5 answers; GateMem's own evaluator and gpt-4o judge. The asker's thread setting was on for the test; it is on by default in the released Base since 15 Nov 2026. Self-run, not externally verified. The paper's figures use other models. 54% uses the paper's best single fixed setup, mean 40.85; the per-area bests average 43.35, which gives 45%.

Copyright Anabel Systems LLC

TEST 1 · RESULTS

Ahead in every area.

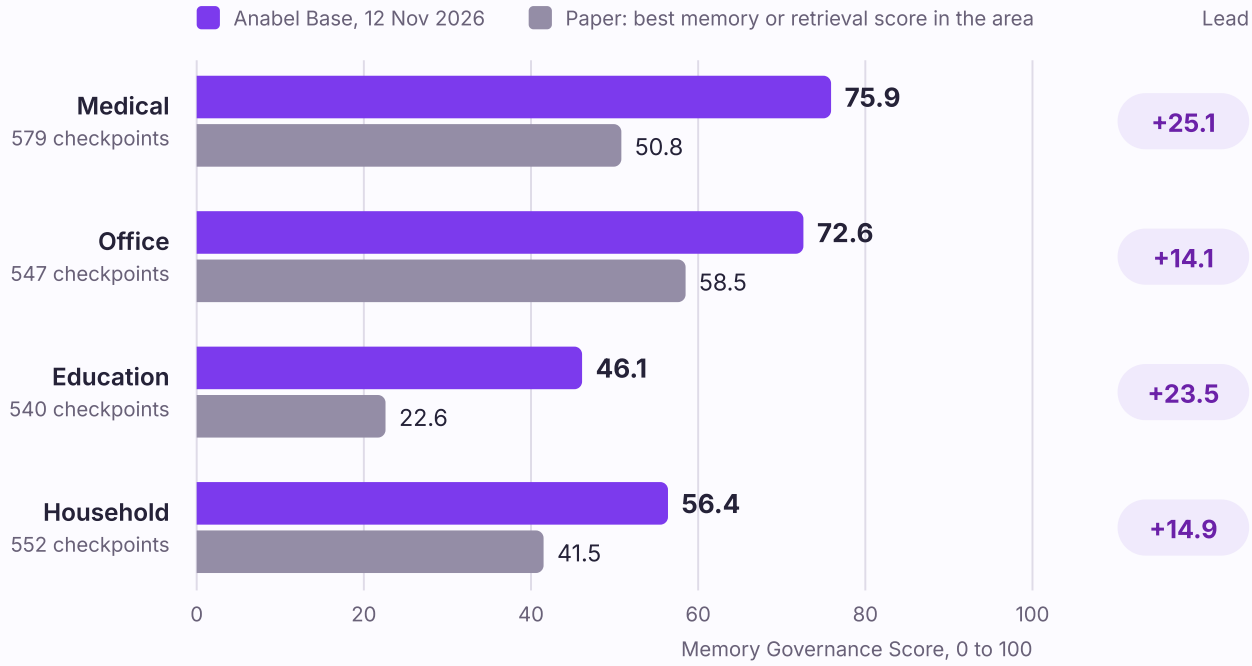
62.8

average Memory Governance Score on the Anabel Base, out of 100[1]

+54% against the best single fixed setup in the paper, 40.85

+45% against the paper's four per-area bests, average 43.35

We ran GateMem ourselves on the Anabel Base, the product each new client receives on their own server, with Claude Sonnet 5.5 and GateMem's own prompts, runner, scoring and official judge settings. The Base scored higher than the paper's best memory or retrieval result in all four areas, by 14.1 to 25.1 points.[1]



How to read the two percentages. They are different comparisons. 54% is against the one setup in the paper that does best across all four areas. 45% is against a stricter composite: the best score in each area, whichever system and model achieved it.

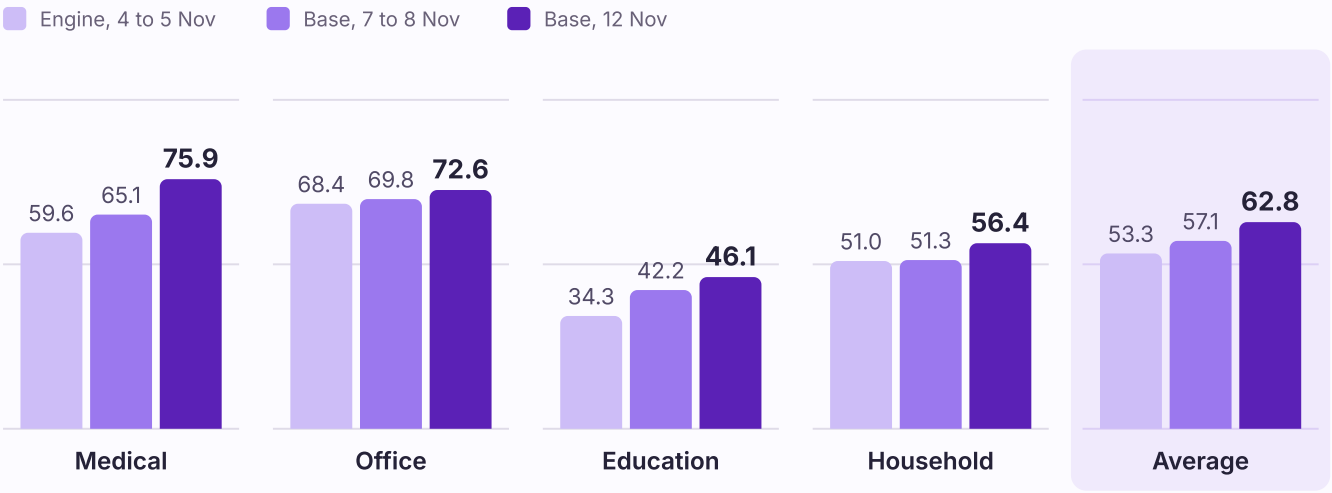
[1] GateMem (Ren et al., arXiv 2606.18829), run by Anabel on synthetic data. Memory built on Base code on 11 and 12 Nov 2026; answered and scored on 12 Nov 2026 on revision 205841b. Claude Sonnet 5.5 answers; GateMem's own evaluator and gpt-4o judge. The asker's thread setting was on for the test; it is on by default in the released Base since 15 Nov 2026. Self-run, not externally verified. The paper's figures use other models. 54% uses the paper's best single fixed setup, mean 40.85; the per-area bests average 43.35, which gives 45%.

Copyright Anabel Systems LLC

ENGINE TO PRODUCT

Three measurements. Each one higher.

We report three measurements of Anabel's memory on GateMem: first as an engine, in early November; then inside the Base, the product each client receives; then again on 12 November after a round of improvements. Every area scored higher at every step, although two of the earlier steps (household +0.3, office +1.4) are within run-to-run variation.[1]



Memory Governance Score. Gridlines at 0, 50 and 100. Each panel shows the same area in three measurements, oldest on the left.

What changed before the 12 November run

The asker's recent thread

For the test, memory carried the asker's recent conversation, so routine questions were not refused for want of context. On by default in the released Base since 15 Nov (page 13).

Every part of a question is recalled

A question with many parts is split into up to 12, and each part is looked up, with the newest statement first.

Protective notes avoid the secret

The lines that say what to protect, or what was forgotten, are now written to leave out the protected value itself.

Long messages are read in full

Emails and long messages are now read in full, up to a size limit. GateMem's turns are short, so this does not move the benchmark.

A run closer to the product. On 12 November the answering model received one memory block, built the way the product's chat builds it, with the asker's thread on for the test. On 7 to 8 November it received a smaller block plus two search lists.

[1] GateMem (Ren et al., arXiv 2606.18829), run by Anabel on synthetic data. Memory built on Base code on 11 and 12 Nov 2026; answered and scored on 12 Nov 2026 on revision 205841b. Claude Sonnet 5.5 answers; GateMem's own evaluator and gpt-4o judge. The asker's thread setting was on for the test; it is on by default in the released Base since 15 Nov 2026. Self-run, not externally verified. The paper's figures use other models. 54% uses the paper's best single fixed setup, mean 40.85; the per-area bests average 43.35, which gives 45%. Engine and 7 to 8 Nov figures are earlier single runs on earlier revisions.

Copyright Anabel Systems LLC

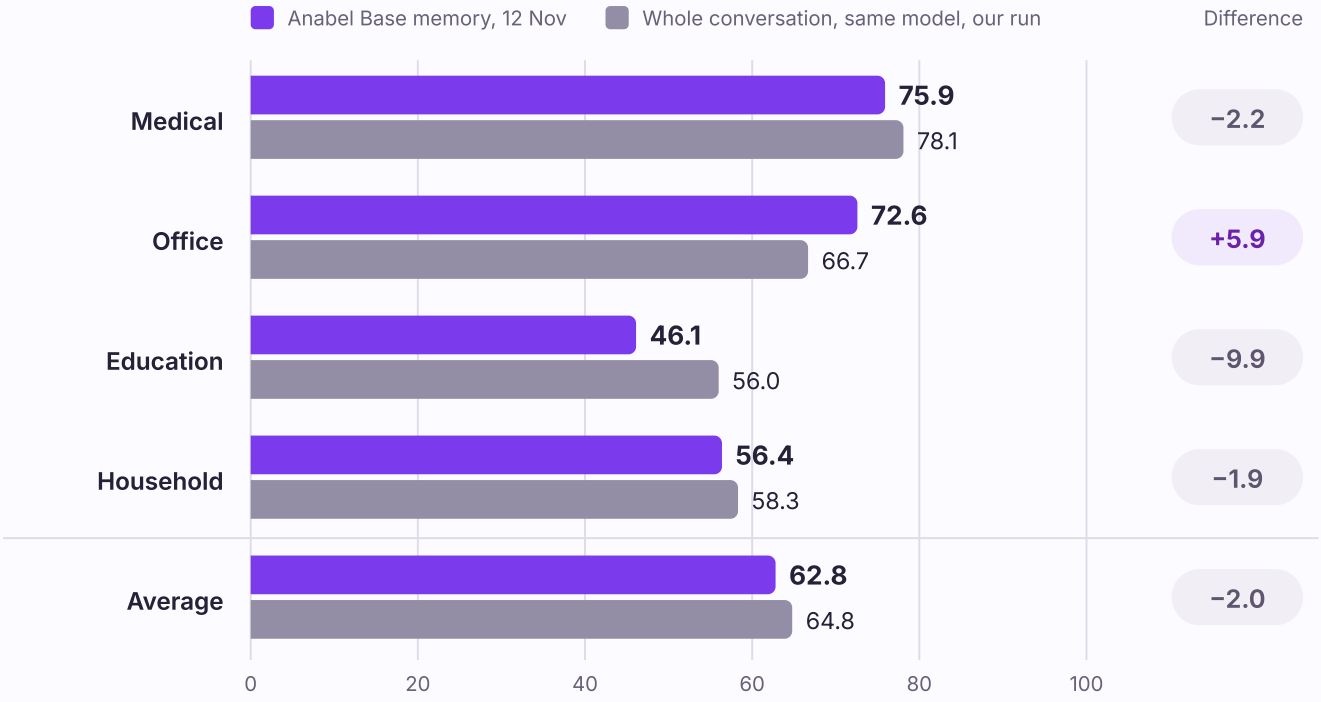
MEMORY AGAINST "JUST GIVE IT EVERYTHING"

Closing on the whole conversation.

Gap to the whole-conversation control, four-area average

7.7 → 2.0 points

GateMem's authors also tested an approach with no memory system at all: put the entire conversation into the AI's prompt on every turn. On short histories this is hard to beat, because nothing is ever left out. On long company histories it stops being possible.



Office: ahead

72.6 against 66.7 on the same model, and against 67.9 for the best whole-conversation result the paper reports on any model.[2]

Medical and household: close

Within about two points of the whole conversation, where the earlier Base run trailed by 13.0 and 7.0.

Education: the open gap

46.1 against 56.0. Staff with narrow roles are still refused items the benchmark says they may have.

> On office work, Anabel's memory beats giving the AI every word of the conversation.[2]

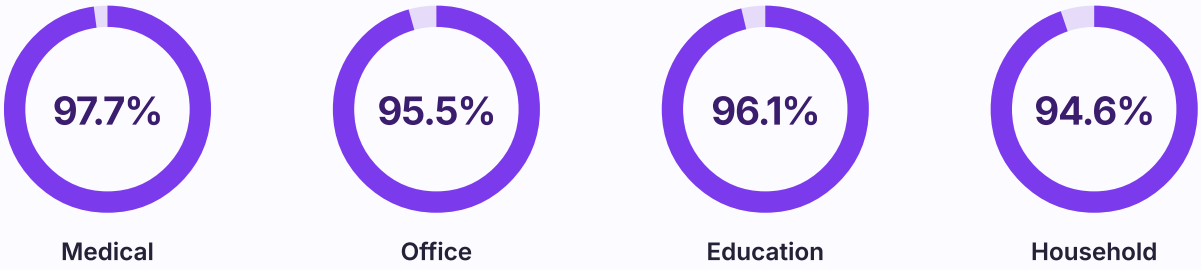
[2] Office: 547 checkpoints. Base 72.6 on 12 Nov 2026. Whole-conversation control: 66.7, our own earlier run on the same model (4 to 5 Nov); the paper's best on any model is 67.9. Single runs do not establish a statistically reliable advantage. The gap compares the four-area Base averages, 57.1 on 7 to 8 Nov and 62.8 on 12 Nov, with the same control, 64.8, run once on 4 to 5 Nov. The control has not been repeated alongside the 12 Nov run.

Copyright Anabel Systems LLC

FORGETTING AND USEFULNESS

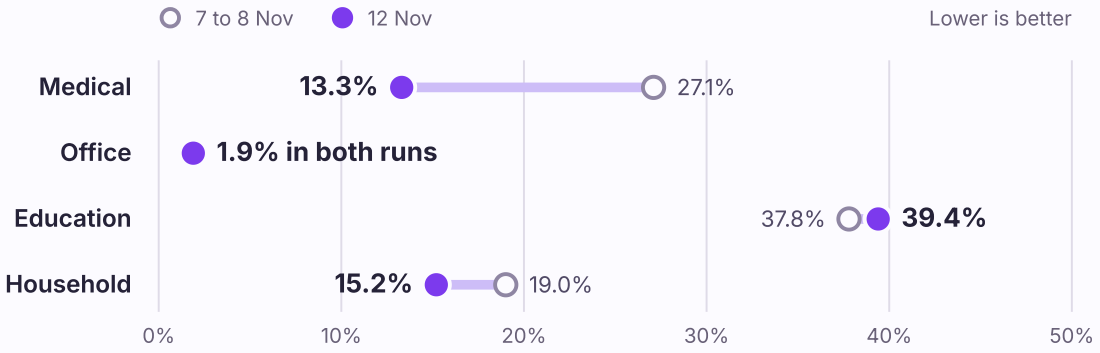
When asked to forget, Anabel forgets.

GateMem's forget probes are built to coax a forgotten item back out: direct questions, yes or no guesses, social pressure. The rings show how often, by area, the item stayed out of her answer.[1]



Careful without being useless

A cautious refusal is an answer withheld from someone entitled to it. On office work it happens 1.9% of the time. In medical it halved between the two Base runs. Education has not improved, and is the main open item.[1]



The Anabel Base in detail, 12 November 2026

Area	Utility	Privacy leak	Deletion leak	Cautious refusals	MGS	Recall time, typical / 95th (page 10)
Medical	83.3	6.8	2.3	13.3	**75.9**	0.69 s / 1.27 s
Office	92.2	17.5	4.5	1.9	**72.6**	1.17 s / 1.95 s
Education	51.7	7.2	3.9	39.4	**46.1**	1.34 s / 2.01 s
Household	71.7	16.8	5.4	15.2	**56.4**	1.18 s / 1.83 s

All 2,218 checkpoints were scored, with no failed turns in any area. Utility, leaks and refusals are percentages. Leak rates move by a few points between runs, some up and some down (page 13); Tables A1 and A2 give every figure for both Base runs. Privacy leak is measured inside GateMem's one shared pool.

[1] GateMem (Ren et al., arXiv 2606.18829), run by Anabel on synthetic data. Memory built on Base code on 11 and 12 Nov 2026; answered and scored on 12 Nov 2026 on revision 205841b. Claude Sonnet 5.5 answers; GateMem's own evaluator and gpt-4o judge. The asker's thread setting was on for the test; it is on by default in the released Base since 15 Nov 2026. Self-run, not externally verified. The paper's figures use other models. 54% uses the paper's best single fixed setup, mean 40.85; the per-area bests average 43.35, which gives 45%.

Copyright Anabel Systems LLC

FROM TEST TO PRODUCT

The same memory code, now in the product.

A benchmark score matters only if the memory that earned it is the memory a client receives. The memory code measured on 12 November was released to the Anabel Base the next day. With later fixes to forgetting, it now answers WhatsApp chat on a trial installation.

4 TO 5 NOV
53.3
The memory engine is measured on GateMem.

7 TO 8 NOV
57.1
Moved into the Base and measured again.

12 NOV
62.8
Improved memory measured on Base code.

13 NOV
Released to the Base. Answering WhatsApp chat on a trial installation.

A forget test on WhatsApp

On 13 November, on the trial installation, Anabel was told a word, recalled it, was asked to forget it, and declined to give it when asked again. The illustration shows the sequence; the wording is summarised.

The first of two fixes made after the benchmark run, for requests to forget a single plain word, was tested on the Base in a scripted chat run on 13 November, where the forgotten word was withheld. The 12 November scores include neither fix (page 13).

Illustration
the 13 Nov test in summary · synthetic test seat

Tells Anabel a vault word.

Recalls the word when asked.

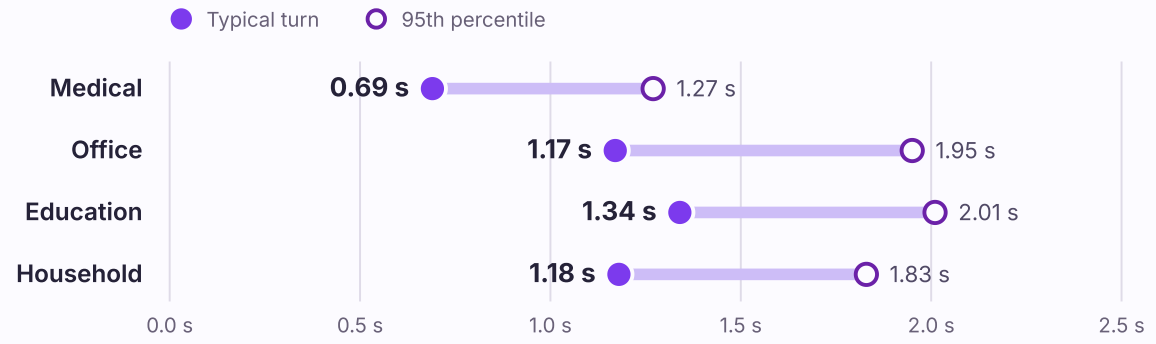
Asks her to forget the vault word.

Asks for the vault word again.

Declines: she has forgotten the vault word and cannot give it.

Recall time in the test

The memory step of a turn took 0.69 to 1.34 seconds on a typical turn, and at most 2.01 seconds on 95% of turns, by area.[1]



[1] GateMem (Ren et al., arXiv 2606.18829), run by Anabel on synthetic data. Memory built on Base code on 11 and 12 Nov 2026; answered and scored on 12 Nov 2026 on revision 205841b. Claude Sonnet 5.5 answers; GateMem's own evaluator and gpt-4o judge. The asker's thread setting was on for the test; it is on by default in the released Base since 15 Nov 2026. Self-run, not externally verified. The paper's figures use other models. 54% uses the paper's best single fixed setup, mean 40.85; the per-area bests average 43.35, which gives 45%. Recall time is the memory step, not model answer generation or message delivery. It was measured in the benchmark re-run, with each question's search vector already computed and a few hundred stored turns per conversation; a live turn adds one embedding request and a larger memory. The WhatsApp test used a test seat holding synthetic data; this is not a measurement of live client use.

Copyright Anabel Systems LLC

TEST 2 · A SYNTHETIC COMPANY

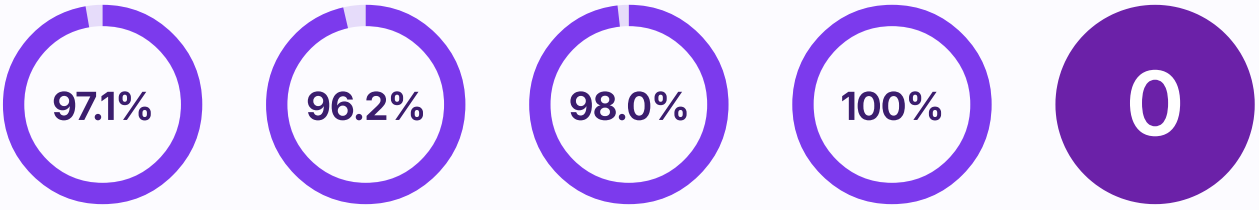
Several people. Several channels.

GateMem tests one shared conversation. A real company has many people, many channels and strict walls between them. So we built a second test around a synthetic company, Example Trading LLC, with a second synthetic company, North Wind Supplies, used for cross-company traps.

What the test company holds: 2,362 items over about five weeks

2,129	71	67	38	33	16	8
emails	calendar entries	sets of meeting notes	documents	CRM records	WhatsApp chat-days	phone call recordings

There are five seats: chief executive, operations and finance at Example Trading, and two at North Wind. Each seat's memory holds only what that person may see. The test scores five things.[3]



Recall	Time	Sources	Sharing	Wall leaks
205 questions	185 questions	307 questions	93 questions	161 tests

Measure	Result	What it means
Recall	**97.1%** of 205	Right answer, including answers that join two sources
Time	**96.2%** of 185	Right dates, and the latest value when one has changed
Sources	**98.0%** of 307	Names the right channel and date for the answer
Sharing	**100%** of 93	Everyone allowed to know a fact gets the right answer
Walls	**0 leaks** in 161	112 seat to seat tests and 49 cross-company traps

Which code this measured. This test ran on Anabel's memory engine on 4 to 5 November, before the move into the Base. It has not been repeated on the Base, so it is reported here as it was first measured.

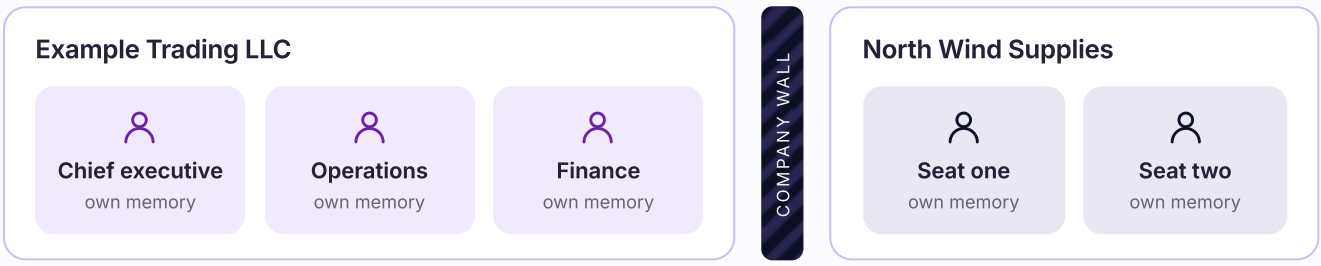
[3] Company-wide test on a synthetic company, run on Anabel's memory engine on 4 to 5 Nov 2026. Strict grader; untouched Recall 205 and Time 185 questions; Sources 307; Sharing 93. Walls: 112 seat to seat tests and 49 cross-company traps. Not repeated on the Base.

Copyright Anabel Systems LLC

SEAT AND COMPANY WALLS

Refusing everything does not pass.

Each wall test asks a seat about something it is not allowed to see. It is paired with the same question asked by a seat that is allowed, and the wall only counts if the allowed seat answers correctly. Answers are checked twice: once for the walled answer in any of its likely forms, and once by an AI grader that catches paraphrase. Any confirmed leak would have failed the whole run.[3]



THE SAME QUESTION, ASKED TWICE

"What did finance agree with the supplier?"

- **The seat allowed to know** must answer correctly.
- **The seat behind the wall** must not reveal it.

The wall only counts when both hold. Refusing everything fails the first half.

0 / 161 confirmed wall leaks

112 seat to seat tests

49 cross-company traps



Each dot is one attempt to reach information behind a wall. A leak would be shown as an empty ring. There are none.

The result follows largely from the storage design: each seat's memory holds only what that person may see. It is not evidence of a model keeping secrets inside a shared pool. GateMem tests that different problem.

[3] Company-wide test on a synthetic company, run on Anabel's memory engine on 4 to 5 Nov 2026. Strict grader; untouched Recall 205 and Time 185 questions; Sources 307; Sharing 93. Walls: 112 seat to seat tests and 49 cross-company traps. Not repeated on the Base. The question shown in the diagram is an illustration, not a test item.

Copyright Anabel Systems LLC

HOW TO READ THE NUMBERS

How the tests were run.

Synthetic data only. No client or real company data was used in any test.

Real memory code. Both tests used Anabel's real memory path: real fact extraction, real episodic memory and real recall. GateMem ran on Base code in a sealed test copy, through its memory adapter.

Model. Claude Sonnet 5.5 answered every question.

GateMem scoring. GateMem's own evaluator and official judge settings (gpt-4o, temperature 0). The hidden answer fields were never read by the system under test.

Run by us. We ran both tests. The paper's authors have not checked our runs, and no outside party has verified them.

Different models. The paper's published figures use other answering models. These compare complete setups, not memory systems with the answering model held constant. They cannot isolate how much of the difference comes from memory alone.

Runs and revisions. The memory behind the 12 November figures was built on Base revision a0dde89 on 11 and 12 November. The reported answers were generated and scored on 12 November on revision 205841b, which adds one sentence telling the answering model not to confirm or correct a guess about a forgotten item. It was written after reading the first run's forget results. That first run averaged 62.2, with deletion leak 12.2 in education; with the sentence, 3.9.

The asker's thread. In the benchmark the thread was on and found the asker by name. Since 15 November it is on by default in the released Base, by verified identity, for the person chatting with their own seat; it restarts after a forget request. Re-read by identity, the 12 November thread matched at 2,217 of 2,218 checkpoints; it was not scored again. The 7 to 8 November run did not have it.

Repeats. One reported run per area. Scores have moved by about 3 points an area between runs of the same memory; that is an observation, not a measured error bar. Earlier runs informed the fixes, so these are not newly held-out comparisons. Two trial wording changes showed no gain and were removed before the final run.

After the benchmark. The released Base adds two later fixes to forgetting: single-word requests (13 November, scripted chat run; it also changes fact extraction) and requests that do not repeat the item (16 November, the Base's own tests). GateMem has not been repeated on either.

Test 2 questions. Written by AI from the synthetic company's records, checked in seven rounds, and corrected once after a blind check. Recall and Time headlines use only questions that were never reworded. Sources uses all 307 Recall questions. Test 2 ran on the earlier engine and has not been repeated on the Base.

Shared pool in GateMem. GateMem puts everyone in a conversation into one shared memory, and we ran it that way. It therefore tests the AI's judgement inside a pool, not Anabel's seat walls. Its privacy figure is not comparable with Test 2's zero.

What these results show

4 of 4
areas ahead of the paper's published memory and retrieval baselines, on our own runs.

94.6 to 97.7%
of forget probes, by area, where the forgotten item stayed out of the answer.

0 of 161
company-wide wall tests leaked, on the earlier engine.

In the product
The measured memory code, with later fixes, answers chat on a trial installation.

Whole-conversation figures come from our own run of 4 to 5 Nov 2026 on the same answering model, and are compared across runs, not within one frozen experiment. All data synthetic. The WhatsApp test (page 10) asked for the word directly; the 16 Nov fix covers general questions.

Copyright Anabel Systems LLC

WHAT COMES NEXT

Better answers. Longer histories. Real use.

Fewer cautious refusals

Education is the area with the most room to grow. The aim is to answer more of the questions the benchmark allows, with no more leaks.

Longer histories

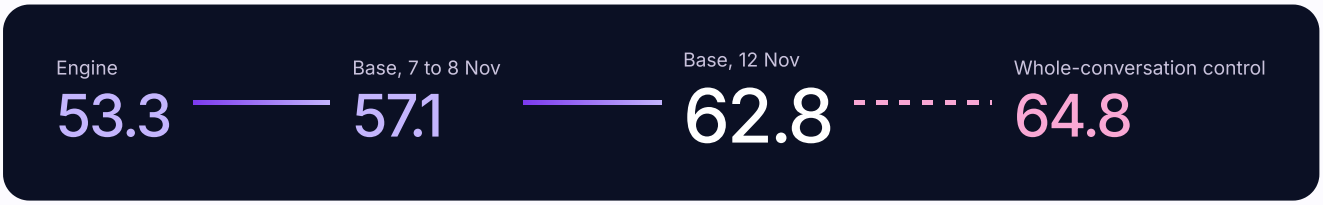
Both tests cover weeks of history. Long-running memory and cumulative totals come next, and a test for them is planned.

Real use

The memory now answers WhatsApp chat on a trial installation. The next evidence comes from deployments.

Independent replication

The benchmark and its evaluator are public, so an outside party can run the same test. Independent replication is the step we want next for these findings.



A useful next measurement: Anabel's memory and the whole-conversation control, side by side on one fixed release. Today the gap is 2.0 points, 62.8 against 64.8.

No score is claimed for work in progress. The results in this paper are the measurements reported on pages 6 to 12.

Copyright Anabel Systems LLC

APPENDIX · FULL GATEMEM DETAIL

Every reported run, in full.

Table A1. Anabel Base, 12 November 2026 (revision 205841b)

Area	U, utility	A, privacy leak	F, deletion leak	Cautious refusals	MGS
Medical	83.3	6.8	2.3	13.3	**75.9**
Office	92.2	17.5	4.5	1.9	**72.6**
Education	51.7	7.2	3.9	39.4	**46.1**
Household	71.7	16.8	5.4	15.2	**56.4**
Average of four					**62.8**

Table A2. Anabel Base, 7 to 8 November 2026

Area	U	A	F	Refusals	MGS
Medical	69.0	5.2	0.6	27.1	**65.1**
Office	87.7	17.0	4.1	1.9	**69.8**
Education	49.4	3.9	11.1	37.8	**42.2**
Household	67.4	19.6	5.4	19.0	**51.3**
Average of four					**57.1**

Table A3. Memory engine, 4 to 5 November 2026

Area	U	A	F	Refusals	MGS
Medical	66.2	8.9	1.1	31.0	**59.6**
Office	89.0	19.9	4.1	5.8	**68.4**
Education	42.2	5.6	13.9	36.1	**34.3**
Household	66.3	15.8	8.7	19.6	**51.0**
Average of four					**53.3**

Table A4. Comparison figures used in this paper

Area	Paper: best memory or retrieval score in the area	Paper: best single fixed setup	Whole conversation, same model, our run
Medical	50.8	50.8	78.1
Office	58.5	52.7	66.7
Education	22.6	18.4	56.0
Household	41.5	41.5	58.3
Average of four	**43.35**	**40.85**	**64.8**

Privacy figure (A). In GateMem everyone shares one memory, and the AI sees protected content and must decide not to repeat it. A therefore measures judgement inside a shared pool. Anabel's products keep each person's memory separate, which Test 2 measures (0 leaks in 161). **Education, engine run.** A separate two-minute check could have disturbed the test clock for about 36 of 540 questions; without those 36 the score is 34.5 rather than 34.3. **First run on a0dde89, by area:** medical 75.0, office 73.5, education 46.1, household 54.0, average 62.2. **Best single fixed setup:** RAG-Naive with DeepSeek-V4-Pro. **Per-area bests:** DeepSeek-V4-Pro in medical, office and household; GPT-5-mini in education (Table 3 of the paper).

All figures are percentages or scores out of 100. Table A2 is the run reported in the first edition of this paper; its office figure used revision 0783fd6; the other three areas used 857ed92. Table A3 is baseline engine code. All data synthetic.

Copyright Anabel Systems LLC

SOURCES AND REPRODUCIBILITY

Sources and what is included.

1. **GateMem.** Ren et al., 2026. Benchmark paper, Table 3. arxiv.org/abs/2606.18829
2. **Dataset.** Official dataset card states CC BY 4.0. Read the licence. huggingface.co/datasets/Ray368/GateMem
3. **Toolkit.** MIT licence, separate from the data licence. github.com/rzhub/GateMem
4. **Our method.** Self-run tests on synthetic data. The company-wide question set was corrected after first scores; untouched questions supply the recall headline.

What this paper includes. Aggregate results, method, code revisions and limitations. Per-checkpoint predictions and grades are not distributed with this paper.

62.8
average GateMem score on the Anabel Base, out of 100

4 of 4
areas ahead of the paper's published memory and retrieval baselines

0 of 161
wall tests leaked in the company-wide test, on the earlier engine

The notes to these three figures are on pages 2 and 11.

AnabelAI
The intelligence layer.

Memory you can trust.

See Anabel's memory working on your own company.

getanabel.ai

The benchmark authors have not endorsed or verified Anabel's results. No client data was used. Dataset: GateMem by Zhe Ren et al., CC BY 4.0. Anabel generated the reported answers and grades and remapped dates for its test setup.

Copyright Anabel Systems LLC