

# Facts Without Owners. Consolidating a Conversation into a Language Model's Weights Stores What Was Said, Not Who Said It

Maximiliano Rodrigo Speranza

[maximiliano.speranza@gmail.com](mailto:maximiliano.speranza@gmail.com)

Independent Researcher <https://orcid.org/0009-0005-0413-8554>

---

## Research Article

**Keywords:** conversational memory, knowledge consolidation, parametric memory, attribution, abstention, LoRA, LoCoMo, pre-registration

**Posted Date:** October 20th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11394470/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** The authors declare no competing interests.

---

# Facts Without Owners

Consolidating a conversation into a language model’s weights stores what was said, not who said it

Maximiliano Speranza

Independent Researcher, Buenos Aires, Argentina

[ORCID 0009-0005-0413-8554](https://orcid.org/0009-0005-0413-8554)

[maximiliano.speranza@gmail.com](mailto:maximiliano.speranza@gmail.com)

October 2026

## Abstract

A personal assistant that learns from its user could keep the conversation in its weights instead of in an external archive. We consolidate three real-length two-person conversations from LoCoMo into Qwen3-4B with a LoRA adapter trained on questions that the same model writes about the raw turns, and we test it on the benchmark questions, which it never saw. Consolidation works. Without any archive, accuracy goes from 0.00 to 0.28–0.39 across the three conversations, a repeated training run gives the same result, and we detect no loss on TriviaQA. But the model does not store who said each fact. Asked the same question about the other speaker, it answers instead of abstaining 67–91 % of the time. In about three of every ten such questions, its answer matches a first-person turn of the other speaker. We then ran seven pre-registered variants that use only the weights or the model’s own signals, namely inverse questions, “I don’t know” examples with the speaker swapped at several doses, and a threshold on the model’s confidence. On the two conversations held out from development, every variant that reaches 0.40 abstention on these questions loses 0.09–0.17 of accuracy, and the one variant that looked free on the development conversation did not replicate. A model-free check against the raw archive does better. It asks whether the answer comes from a turn of the other speaker. It raises abstention from 0.17 to 0.45–0.47 at a cost of 0.005–0.042 in accuracy, between a fifth and a half of the cost of the weight-only variants at the same abstention. It still misses our pre-registered cost limit in one of two runs (7.0 % of correct answers lost, limit 5 %). The practical conclusion is that consolidation into weights should keep the archive, at least as the source of provenance.

## 1 Introduction

An assistant used every day accumulates a long history with its user. There are two ways to make it remember. The external way keeps the history in an archive and retrieves what each question needs (Packer et al., 2023; Wu et al., 2025). The internal way writes the history into the weights, as a person does when a conversation becomes knowledge (Dennis et al., 2026; Li, 2026; O’Neill et al., 2026; Padmanabhan et al., 2026). The internal way promises that knowledge stays available when retrieval fails, and that it can be combined with what the model already knows. It also has known risks. Fine-tuning on new facts can encourage hallucination (Gekhman et al., 2024), and facts learned in one direction are not always available in the other (Berglund et al., 2023).

We ask a narrow question. When a real-length conversation between two people is consolidated into the weights, *what is stored and what is lost?* LoCoMo (Maharana et al., 2024) gives a sharp test for this, because its adversarial category asks about one speaker something that only the other speaker said. The correct answer is to abstain. We call these *swapped-person questions*.

## Contributions.

1. Consolidating the raw turns of a conversation into a 4B model with a self-generated question set works. Without an archive, the model answers 28–39 % of the benchmark questions on three conversations, with the same result in a repeated run and no detectable loss on TriviaQA (Section 3.1).
2. The dominant failure is attribution. The model knows the fact but not its owner. It abstains on only 9–33 % of the swapped-person questions, and in about 30 % of them its answer matches a first-person turn of the other speaker (Section 3.2).
3. Seven pre-registered variants use only the weights or the model’s own confidence. On the held-out conversations, all of them buy abstention with accuracy at a similar rate. The trained “I don’t know” learns the form of the trap, not what the model knows (Section 3.3).
4. A model-free provenance check against the raw archive buys the same abstention for a fifth to a half of the cost (Section 3.4). It did not meet our pre-registered cost limit in one of two runs, and we report it as not adopted under our own rule.

Every experiment was pre-registered in a time-stamped repository before it was run. Every evaluator was tested on synthetic data before it saw real outputs, and every label was committed before any number was computed (Section 2).

## 2 Setup

**Data.** We use three LoCoMo conversations (Maharana et al., 2024), conv-42, conv-43 and conv-48, each between two people, with 29–30 sessions and 629–681 turns. Together they have 568 questions in categories 1–4 (multi-hop, temporal, open-domain and single-hop) and 173 swapped-person questions (category 5). The consolidation never sees any benchmark question.

**Model and consolidation.** The model is Qwen3-4B (Yang et al., 2025). The base model (*version 0*) reads the conversation in windows of up to ten turns, with the session date, and writes up to six short-answer questions per window, which it then rephrases in two ways. In 15 % of the questions about one speaker, the name is swapped for the other speaker’s and the target becomes “Not mentioned in the conversation.”, which is how LoCoMo builds its traps. This gives 1,246 to 1,334 examples per conversation. A rank-16 LoRA adapter (Hu et al., 2022) is trained on the seven projections of all 36 layers for three epochs, with batch 8, learning rate  $2 \cdot 10^{-4}$  with a 20-step warm-up and cosine decay, and loss only on the answer. A KL term (weight 1.0) ties it to version 0 on two anchor texts per step, SQuAD paragraphs and chat requests answered by version 0. One run, including writing the questions, training and the exam, takes 38 to 45 minutes on a T4 GPU. We call this recipe D1.

**Exam.** Every benchmark question is answered greedily, up to 24 tokens, in four arms.

- **B0**, version 0 without the archive;
- **D**, the consolidated model without the archive;
- **B0S**, version 0 with archive turns;
- **DS**, the consolidated model with archive turns.

The archive turns are retrieved by TF-IDF. They are the ten best turns for the question, each with its two neighbours. General knowledge is measured on 200 TriviaQA questions (Joshi et al., 2017).

**Labels.** Answers to categories 1–4 were labelled correct or incorrect against the LoCoMo reference. The labeller was an LLM, Claude (Anthropic), following a written rule fixed before the first labelling. It saw the question, the reference and the answer, but not the arm, the recipe or any confidence score. Borderline cases were marked as doubtful and counted as correct. Two examples are a relative date that the session date resolves, and a list with one extra item. Decisions use the labels as given. A strict variant counts the doubtful labels as wrong. It was reported with each pre-registered result. It changes two results. It weakens the gain of consolidation with the archive on conv-42 (Section 3.1), and it brings the archive check within its cost limit (Section 3.4). Identical answers to the same question share their label. Where the same answer was labelled on two different sheets (514 answers on conv-42 and 599 on conv-43), the labels agreed in 98.1 % and 100 % of the cases. For the 10 disagreements we keep the label of the earlier sheet. An abstention counts as wrong on categories 1–4. On category 5, only abstention is correct, and it is detected by string match. Accuracy is the fraction of category 1–4 questions answered correctly. Intervals are 95 % percentile bootstrap intervals over questions, with 10,000 resamples.

## 3 Results

### 3.1 The conversation enters the weights

Version 0 abstains on 99.5–100 % of the questions without the archive, so its accuracy is 0.000. The consolidated model answers from memory (Table 1). We trained the recipe twice with the same data and seed. The two runs differ only through non-deterministic GPU computation, and they agree within 0.016 on every conversation. In run 2, TriviaQA accuracy is 0.335–0.355 for the consolidated models and 0.340 for version 0, so no damage is detected at  $n = 200$ . With the archive, consolidation still adds accuracy. DS minus BOS is +0.055 on conv-42 (95 % CI +0.005 / +0.106, pre-registered; −0.005 under strict labels), and +0.04 and +0.07 on the other two (exploratory). Dates do not consolidate. On temporal questions the model reaches only 0.08 with the three conversations pooled.

**Table 1.** Accuracy without archive (categories 1–4) and abstention on swapped-person questions, for two runs of the consolidation recipe with the same data and seed. Version 0 scores 0.000 accuracy and abstains on every swapped-person question.

|       | conv-42 ( $n$ 199 / 61) |            | conv-43 ( $n$ 178 / 64) |            | conv-48 ( $n$ 191 / 48) |            |
|-------|-------------------------|------------|-------------------------|------------|-------------------------|------------|
|       | accuracy                | abstention | accuracy                | abstention | accuracy                | abstention |
| run 1 | 0.286                   | 0.131      | 0.393                   | 0.094      | 0.283                   | 0.333      |
| run 2 | 0.302                   | 0.164      | 0.388                   | 0.094      | 0.298                   | 0.312      |

### 3.2 But not who said it

On swapped-person questions the consolidated model abstains only 9–33 % of the time (Table 1). A typical case is “What color did Joanna choose for her hair?”, answered “Purple”. It was Nate who dyed his hair purple, and the conversation never mentions the color of Joanna’s hair. The model has the fact. What it lacks is the owner. Section 3.4 measures this directly. In 0.27–0.30 of the swapped-person questions, the model’s answer from memory shares enough words with a first-person turn of the *other* speaker for a lexical check to find it, and that turn scores at least 1.5 times higher than any turn of the person asked about.

### 3.3 Fixing it with the weights alone

Each of the following was pre-registered before it was run. The variants meant as fixes (D2, D3, D4a, D4b, the “I don’t know” cell of D5 and D6) had the same two pass criteria. They needed at least 0.50 abstention on swapped-person questions (D2–D4 also +0.25 over D1), and an accuracy loss whose 95 % interval stays above  $-0.10$ .

- **D2** adds inverse questions, of the form “Who dyed their hair purple?”  $\rightarrow$  “Nate”, as a remedy for the reversal curse (Berglund et al., 2023), plus swapped-person “I don’t know” examples for every eligible question.
- **D3** is D2 without the “I don’t know” examples whose swapped question coincides with a true question about the other speaker.
- **D4a** and **D4b** use the D3 recipe at half and a quarter of the “I don’t know” dose.
- **D5** is a  $2 \times 2$  design that crosses inverse questions with the “I don’t know” dose. Its inverse-only cell is a diagnostic, not a fix.
- **D6** keeps the D1 weights and adds abstention at inference when the mean token log-probability of the greedy answer falls below a threshold (Kadavath et al., 2022). The threshold was fixed on conv-42 and tested on conv-43 and conv-48.

On the held-out conversations none passed. D2 passed on conv-42, where these recipes were first tried, with 0.54 abstention and no loss of accuracy, but it failed its pre-registered replication on conv-43. Figure 1 shows why. On the held-out conversations all variants sit on one trade-off line, and each point that reaches 0.40 abstention or more loses between 0.09 and 0.17 of accuracy. Two diagnostics explain the line.

*The trained “I don’t know” learns the form of the trap, not the model’s knowledge.* On conv-43, recipe D3 abstains on 0.30 of the questions that D1 answered correctly and on 0.32 of those it answered wrongly. Its precision when it does answer does not change ( $0.43 \rightarrow 0.42$ ). The inverse questions are not the cost. With few “I don’t know” examples, adding them changes accuracy by  $-0.022$  (CI  $-0.073 / +0.028$ ). The “I don’t know” examples are the cost. Without inverse questions, their full dose changes accuracy by  $-0.163$  (CI  $-0.230 / -0.101$ ).

*The model’s confidence does not see the trap.* On correct answers against swapped-person answers, the AUROC of the mean log-probability is 0.69 on conv-42, but only 0.54 on conv-43 and 0.56 on conv-48. The model answers a question about the wrong person as confidently as one about the right person, because it does know the fact.

### 3.4 Checking provenance against the archive

The archive holds every turn with its speaker. We reuse, unchanged, a model-free check that was developed earlier for an external reader that answers from retrieved turns. Its margin was fixed on that reader’s answers for conv-42 and conv-43. On five held-out conversations, one of them conv-48, it had raised that reader’s abstention on swapped-person questions from 0.47 to 0.66 while losing 1.6 % of its correct answers. The check works as follows.

1. The question must name exactly one speaker  $X$ .
2. The check takes the answer, with the speakers’ names removed, and scores the turns that the archive would retrieve for that question by TF-IDF similarity to it.
3. Let  $s_Y$  be the best score among the turns of the other speaker and  $s_X$  the best among the turns of  $X$ .
4. The answer is replaced by “I don’t know” if  $s_Y > 0$ ,  $s_Y > 1.5 s_X$ , and the best turn of the other speaker is in the first person.

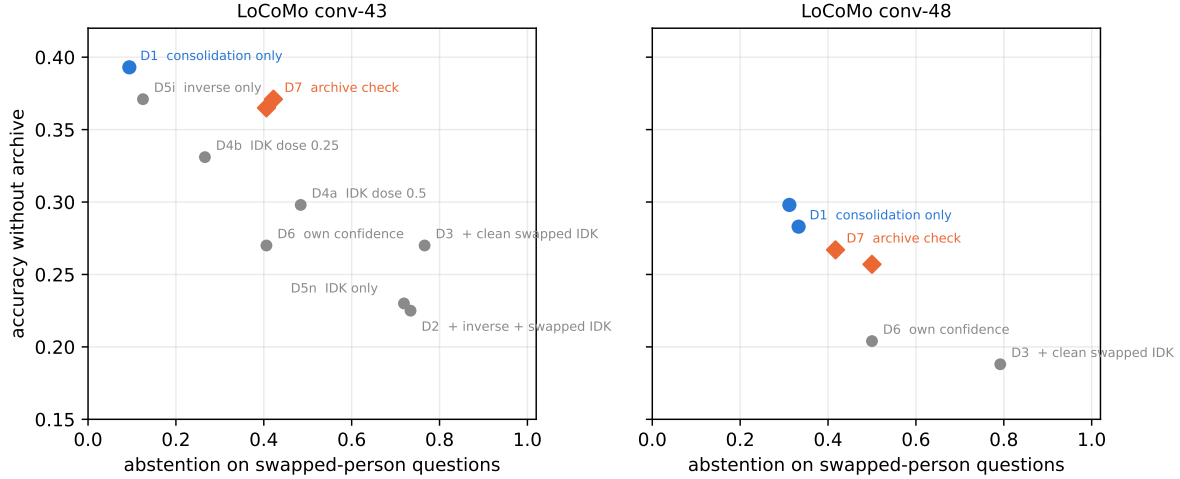


**Figure 1.** Abstention on swapped-person questions against accuracy without archive, for every pre-registered recipe on the two held-out conversations. Grey points use only the weights or the model’s own confidence. The orange diamonds are the archive check of Section 3.4 on the two D1 runs. Blue points are D1 alone. On conv-48 only D1, D3, D6 and D7 were evaluated.

The margin 1.5 was fixed in that earlier study and was not tuned here. The check needs no model and no new labels, because an answer it turns into “I don’t know” simply counts as wrong. The criteria were pre-registered. Abstention on swapped-person questions had to rise by at least +0.15 with a 95 % interval above zero. At most 5 % of the correct answers could be lost. Both had to hold in each of the two runs, with the three conversations pooled.

**Table 2.** The archive check on the answers of the consolidated model without archive. The three conversations are pooled (173 swapped-person questions).

|       | abstention on swapped-person questions | difference (95 % CI)     | correct answers lost |
|-------|----------------------------------------|--------------------------|----------------------|
| run 1 | 0.173 → 0.445                          | +0.272 (+0.208 / +0.341) | 8 / 181 (4.4 %)      |
| run 2 | 0.179 → 0.474                          | +0.295 (+0.231 / +0.364) | 13 / 186 (7.0 %)     |

The gain in abstention passes in both runs (Table 2). The cost limit passes in run 1 and fails in run 2, so the check is *not adopted* under our rule. A first evaluation reported 4.5 % for run 2, because a labelling file was not loaded and 38 answers had no label. We corrected the loader and report the corrected figure. Both outputs are kept. Under strict labels, which are reported but do not decide, both runs stay within the limit, with 3.6 % and 4.2 % lost. Seven of the thirteen correct answers lost in run 2 were doubtful ones. The check removes 17 % of the doubtful correct answers and 4 % of the firm ones, so it mostly removes partial or vague answers.

Per conversation, accuracy changes by  $-0.005$  to  $-0.042$ . At the same abstention, the weight-only recipes lose two to five times more. On conv-43, D6 reaches 0.41 abstention for  $-0.12$ , while the check reaches 0.41–0.42 for  $-0.02$  (Figure 1). Temporal questions gain abstention ( $0.06 \rightarrow 0.21$ – $0.22$ ) with no loss of correct answers. The check cannot help when the model invents an answer that matches no turn, and it cannot help when the evidence was said by  $X$  itself.

With the archive in the context (exploratory), consolidation makes the reader more willing to answer a swapped-person question. Version 0 with the archive abstains on 0.53 of them, and the consolidated model with the archive on only 0.19. The check brings the consolidated model to 0.57. Its accuracy with the archive then stays at 0.63–0.65, against 0.58–0.62 for version 0 with the archive and the check.

## 4 Discussion

**What consolidation stores.** The training questions name their speaker, but the answer can be predicted from the topic words alone. The adapter seems to learn the topic-to-answer association and to treat the name as a weak cue. This is consistent with work showing that a fact needs varied formulations to become extractable (Allen-Zhu and Li, 2024) and that learned associations are directional (Berglund et al., 2023). Remedies that act on the training data, such as inverse questions and swapped-name negatives, change how often the model abstains, but they do not bind facts to people. Remedies that act on the model’s own signal cannot see the error, because the model is right about the fact.

**Implications.** Studies of parametric personal memory (Dennis et al., 2026; Li, 2026) measure how well the model recalls and uses the user’s information. Our results suggest that they should also report *who* the model believes said it, for example with swapped-person questions, since recall can rise while attribution stays poor. Source-aware training (Khalifa et al., 2024) attacks a related problem for pretraining documents. For conversational memory, keeping the raw archive gives a cheap and auditable source of provenance, and it does not need to be retrieved for every answer, only checked.

### Limitations.

- We use three conversations of one benchmark, whose dialogues are machine-generated, and one model size (4B).
- The labels come from one LLM labeller with a fixed rule, not from human annotators.
- Consolidation is a single LoRA recipe, and the archive check is lexical.
- The swapped-person questions of LoCoMo measure one kind of attribution error. Errors where both speakers did similar things are harder and are only partly covered.
- The confidence threshold of D6 was fixed on one conversation (conv-42), which turned out to be the easiest one.

## 5 Conclusion

Consolidating a conversation into a 4B model’s weights stores a substantial part of what was said, without detectable damage. It does not store who said it. Teaching the model to abstain, or trusting its confidence, trades knowledge for abstention at the same poor rate. Checking the answer against the raw archive is far cheaper, although in one of our two runs it was still slightly above the cost limit we had set. Memory in the weights and the archive are complementary. The weights hold the knowledge, and the archive holds its owners.

## Declarations

**Funding.** None. **Competing interests.** The author declares no competing interests. **Contributions.** Sole author, responsible for design, implementation, analysis and writing. **Ethics.** Not applicable. There are no human or animal subjects, and the data come from a public benchmark. **Use of tools.** The experimental code and the writing were done with the assistance of a language model (Claude, by Anthropic), which also produced the blind correctness labels described in Section 2. The design, the decisions and the responsibility for every figure belong to the author. **Availability.** The pre-registrations, the evaluators with their synthetic tests, the blind labels, the model outputs and the code of the recipe are kept in a time-stamped git

repository, and are available from the author on request. **Author note.** This work was carried out without institutional sponsorship, supervision or funding, and is reported in a personal capacity.

## References

- Allen-Zhu, Z., and Li, Y. (2024). Physics of Language Models. Part 3.1, Knowledge Storage and Extraction. ICML 2024. arXiv 2309.14316.
- Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. (2023). The Reversal Curse. LLMs trained on “A is B” fail to learn “B is A”. arXiv 2309.12288.
- Dennis, S., Shabahang, K., Guo, H., and Patil, R. (2026). Beyond Inference-Only Deployment. Comparing Weight-Based Consolidation Against Cascading Compaction. arXiv 2605.24657.
- Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., and Herzig, J. (2024). Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? EMNLP 2024. arXiv 2405.05904.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., et al. (2022). LoRA. Low-Rank Adaptation of Large Language Models. ICLR 2022. arXiv 2106.09685.
- Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). TriviaQA. A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. ACL 2017. arXiv 1705.03551.
- Kadavath, S., et al. (2022). Language Models (Mostly) Know What They Know. arXiv 2207.05221.
- Khalifa, M., Wadden, D., Strubell, E., Lee, H., Wang, L., Beltagy, I., and Peng, H. (2024). Source-Aware Training Enables Knowledge Attribution in Language Models. COLM 2024. arXiv 2404.01019.
- Li, B. (2026). User as Engram. Internalizing Per-User Memory as Local Parametric Edits. arXiv 2606.19172.
- Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., and Fang, Y. (2024). Evaluating Very Long-Term Conversational Memory of LLM Agents. ACL 2024. arXiv 2402.17753.
- O’Neill, C., Kirkby, M., Liu, J., and Psenka, M. (2026). Can a Language Model Learn Facts Continually in Its Weights? arXiv 2607.11020.
- Packer, C., et al. (2023). MemGPT. Towards LLMs as Operating Systems. arXiv 2310.08560.
- Padmanabhan, S., Gul, M. O., and Goyal, T. (2026). Updating Parametric Knowledge with Context Distillation Retains Post-Training Capabilities. arXiv 2602.16093.
- Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., and Yu, D. (2025). LongMemEval. Benchmarking Chat Assistants on Long-Term Interactive Memory. ICLR 2025. arXiv 2410.10813.
- Yang, A., et al. (2025). Qwen3 Technical Report. arXiv 2505.09388.