CRELT explained: why conversation history makes RAG leak more training data

25 November 2026 · 8 min read · blogs.iictp.org/crelt-conversational-rag-training-data-leakage-privacy-explained

Short answer: CRELT (Conversational RAG Extraction Leakage Trajectory) is a framework from an arXiv preprint submitted on 21 November 2026 by Syed Khaderi, Karima Makhlouf and Sami Zhioua that measures, turn by turn, how much real contact data a retrieval-augmented generation system discloses from its training data. The authors report that keeping the full conversation history raised disclosure 3.6 to 4.9 times with retrieval enabled, and that single-turn privacy audits can underestimate leakage. This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors' own claims.

TL;DR The short version

- CRELT measures, turn by turn, how much real contact data a RAG system discloses from its training data.
- Retrieval reduces leakage, but conversation history changes the picture, so single-turn privacy audits can underestimate it.
- Results come from small models and the Enron emails; test multi-turn behaviour and treat conversation-history policy as a privacy control.

Read online, listen to it, or change the text size: https://blogs.iictp.org/crelt-conversational-rag-training-data-leakage-privacy-explained

What is CRELT?

RAG is known to reduce training-data leakage in single-turn use. The paper asks whether that holds across multi-turn conversations. CRELT separates retrieval, conversation history and attack strategy in a full-factorial design, uses a retrieval corpus verified to contain none of the target PII, and counts an item as leaked only if it was absent from the model input at the turn it was produced.

How it works

- **Factorial design.** Retrieval on or off, full or cleared history, different attack strategies.
- **Strict leakage count.** This criterion excluded 29% to 53% of raw matches that were already in the input.
- **Models and data.** GPT-J-6B, Pythia-6.9B and Pythia-12B, targeting real Enron contact details.
- **Control.** A length-matched control with random passages tests whether retrieved content or simply extra context drives the effect.

What is new compared with earlier approaches

Earlier audits mostly test one turn. CRELT tracks leakage along the conversation and filters out items the model was simply given.

Results reported by the source

- Retrieval reduces disclosure under both full and cleared history, consistent with single-turn results.
- A random-passage control reproduces 91% to 99.5% of that reduction under cleared history and 47% to 78% under full history, so much of the benefit comes from added context, not from what is retrieved.
- With retrieval on, keeping full history raises disclosure by 3.6 to 4.9 times; retaining only one previous turn already restores disclosure to about the full-history level.
- Clearing history with retrieval reduces disclosure by 72% to 79%, but does not eliminate it.

Limitations and open questions

- The models are small (6 to 12 billion parameters) and the data is the Enron email set.
- Results for current commercial models may differ.
- The paper tests extraction attacks; other leakage routes are not covered. Sources were read on 25 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to apply it in a privacy review

1. Test multi-turn behaviour, not only a single question.
2. Compare full-history and cleared-history settings.
3. Count only items not already present in the model input.
4. Decide a conversation-history retention policy and document it.

What it means for data-protection and legal teams

Conversation-history policy is a privacy control, not just a product setting. If a chatbot keeps prior turns, a data-protection impact assessment should consider multi-turn extraction. The paper is research on small models; it does not state a legal conclusion. See Gemini Decisions API for another system design that limits free-text generation.

Frequently asked questions

- **What is CRELT?** A framework that tracks how much training-data PII a conversational RAG system leaks, turn by turn.
- **What is the main finding?** With retrieval, full conversation history raised disclosure 3.6 to 4.9 times.
- **Does clearing history help?** It cut disclosure by 72% to 79% but not to zero.
- **Is it peer reviewed?** No, it is a preprint.

Related reading

- Gemini Decisions API
- Robots.txt and AI training controls
- AI Practitioner certification

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Khaderi S, Makhlouf K, Zhioua S. Beyond Single-Turn Privacy: Training-Data Leakage in Conversational RAG. arXiv 2611.28748, 21 Nov 2026 (preprint). arxiv.org/abs/2611.28748

Published by IICTP, an independent technical credential body, on 25 November 2026. The current version of this post is at https://blogs.iictp.org/crelt-conversational-rag-training-data-leakage-privacy-explained. More guides: blogs.iictp.org.