Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Demonstrations Track

RUVA: Personalized Transparent On-Device Graph Reasoning

Gabriele Conte[1] , **Alessio Mattiace**[1] , **Gianni Carmosino**[1] ,
Potito Aghilar[1] , **Giovanni Servedio**[1*] , **Francesco Musicco**[1] ,
Vito Walter Anelli[1*] , **Tommaso Di Noia**[1] and **Francesco Maria Donini**[2]

[1]Politecnico di Bari,
[2]Università degli Studi della Tuscia

Abstract

The Personal AI landscape is currently dominated by "Black Box" Retrieval-Augmented Generation. While standard vector databases offer statistical matching, they suffer from a fundamental lack of accountability: when an AI hallucinates or retrieves sensitive data, the user cannot inspect the cause nor correct the error. Worse, "deleting" a concept from a vector space is mathematically imprecise, leaving behind probabilistic "ghosts" that violate true privacy. We propose **RUVA**, the first "Glass Box" architecture designed for Human-in-the-Loop Memory Curation. RUVA grounds Personal AI in a Personal Knowledge Graph, enabling users to inspect *what* the AI knows and to perform precise redaction of specific facts. By shifting the paradigm from Vector Matching to Graph Reasoning, RUVA ensures the "Right to be Forgotten." Users are the editors of their own lives; RUVA hands them the pen.

1 Introduction

Since their release, Large Language Models have quickly transitioned from novelties to commodities [Liang *et al.*, 2025]. The early industry dream of hyper-personalizing these models via fine-tuning has collapsed under the weight of computational costs and the risk of catastrophic forgetting [Luo *et al.*, 2025]. Consequently, the field has converged on Retrieval-Augmented Generation (RAG) [Lewis *et al.*, 2020] as the only viable path for personalization [Wang *et al.*, 2023; Anelli *et al.*, 2017]. However, the reliance on standard Vector RAG has created a crisis: the "Black Box" problem. Current RAG systems compress a user's life (e.g., emails, photos, chats, calendar entries) into high-dimensional embedding spaces. While efficient for matching semantically similar keywords, the user cannot inspect why the system fails, or fix it [Servedio *et al.*, 2025; Di Palma *et al.*, 2025]. More critically, true privacy requires the ability to delete, but one cannot surgically remove a concept from a vector space; one can only suppress it. This mathematical imprecision leaves behind "probabilistic ghosts", traces of sensitive data that the AI might accidentally resurrect [Wang *et al.*, 2025]. GraphRAG methods [Edge *et al.*, 2024] address these limitations by incorporating graph topology into retrieval.

We introduce RUVA (the Etruscan word for "Brother"), a system designed to be as trustworthy and protective as a close sibling. RUVA is a **neuro-symbolic GraphRAG architecture** [Kautz, 2022] that runs entirely **on-device**, representing a paradigm shift from opaque retrieval to transparent reasoning. Trust begins with sovereignty. By running locally, RUVA ensures that the user's Personal Knowledge Graph (PKG) [Balog and Kenter, 2019; Skjæveland *et al.*, 2024] never leaves their device. There is no cloud to hack; the user's life remains physically in their pocket. Unlike vector stores, RUVA grounds its intelligence in a "Glass Box" architecture. Here, every memory is an explicit node, and every association is a visible edge. This structure highlights a core consideration: *Vectors allow Matching; Graphs allow Reasoning.* Consider the query: *"Did Sarah call before I arrived at work?"* A standard vector system simply retrieves chunks containing "Sarah" and "Work" based on similarity, often hallucinating the temporal relationship. In contrast, RUVA traverses the graph topology, comparing the timestamp of the *Call* node against the *Arrival* node to provide the answer.

Finally, RUVA redefines the user's role from a passive data source to the *Editor-in-Chief*. Because the memory is structured, it is fully editable. If RUVA learns something incorrect or sensitive, the user can use the *"Red Pen"* to perform a precise redaction, cutting that specific node (the reified information object) and its edges. The knowledge is deleted instantly and mathematically, ensuring the "Right to be Forgotten." RUVA is not another chatbot. It is an AI that users can actually trust, because RUVA allows them to see inside its head and edit its memories. The project and the demo video are available at https://ruva-assistant.github.io/.

2 The RUVA Architecture

RUVA is engineered as an offline-first, mobile-native system designed to guarantee data sovereignty. To ensure that personal data never leaves the user's possession, all computation occurs locally on the user's mobile device; the **Personal Knowledge Graph (PKG)** is physically bound to the handset, eliminating cloud-based attack vectors [Aghilar *et al.*, 2023]. Furthermore, the system is optimized for mobile constraints, utilizing quantized models and efficient storage to operate alongside the OS without draining the battery. As

*Corresponding authors.

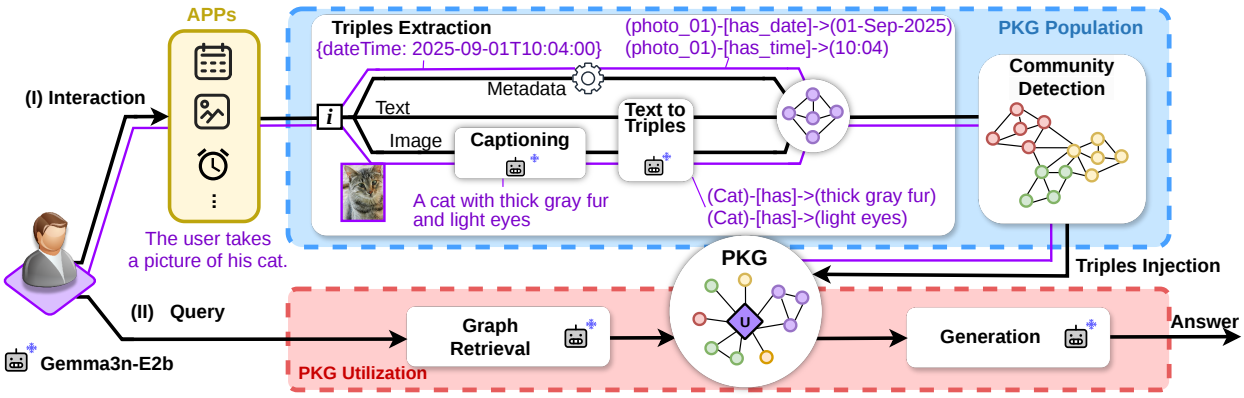


Figure 1: The RUVA architecture. The *Ingestion Workflow* transforms multimodal data into semantic triples to populate the Personal Knowledge Graph. The *Retrieval Workflow* performs graph traversal to generate grounded, hallucination-free answers, entirely on-device.

illustrated in Figure 1, RUVA implements a **Type 3 Neuro-Symbolic architecture** (taxonomy from Kautz [2022]) that separates perception from memory. The *Neural Component* employs a Small Vision Language Model (SVLM, i.e., Gemma-3n-E2b [Kamath et al., 2025]) as the system's "eyes and ears." Its role is to convert unstructured noise (pixels, raw text) into structured signals, accepting the probabilistic nature of recognition. These signals are fed into the *Symbolic Component*, a PKG used for storage and reasoning (Figure 2a). Unlike the neural component, the memory, once created, is deterministic, inspectable, and editable. The system backbone is a hybrid storage engine: a SQLite database extended with `sqlite-vec`. For this research, the engine is extended to accommodate GraphRAG [Edge *et al.*, 2024] operations. This design choice is critical for the "Glass Box" paradigm. The resulting "spiderweb" topology roots at a central `User` node, from which entities (`Events`, `Photos`, `Messages`) branch via typed edges, unifying the user's digital footprint.

2.1 Multimodal Ingestion Workflow

RUVA operates a background "Triple Extraction" pipeline that transforms heterogeneous data streams into semantic triples. For visual data, such as a photo of a train ticket, the image is passed to the local SVLM [Aghilar *et al.*, 2025]. The SVLM generates a dense caption, which is then parsed to extract structured entities, effectively converting pixels into nodes like `(:Receipt)-[:cost]->("95 EUR")`. Textual data, such as emails or notes, is processed directly by the SVLM to extract entities and temporal metadata. To prevent graph fragmentation, RUVA employs an Entity Resolution strategy along with community detection (Leiden algorithm [Traag *et al.*, 2019]). For instance, clustering merges an email from "Sarah Green" and an invite from "S. Green" into a single `Person` node, preventing duplicates.

2.2 Graph-Grounded Retrieval Workflow

Standard Vector RAG fails at complex reasoning because it relies solely on semantic similarity. RUVA solves this via a **multi-step Graph-Grounded Retrieval mechanism**. Upon receiving a query like *"Did Sarah call before I arrived at work?"*, the system first executes an **Anchor Search** using vector similarity to find the relevant "Anchor Nodes" (e.g., the `Person:Sarah` node and the `Location:Work` node). It then performs a configurable **Topological Expansion**, executing an N-hop traversal from these anchors to identify connected events. Finally, during **Answer Generation** (Figure 2b), the retrieved subgraph, containing explicit timestamps and relationships, is serialized and fed into the SVLM.

2.3 The Deletion Mechanism

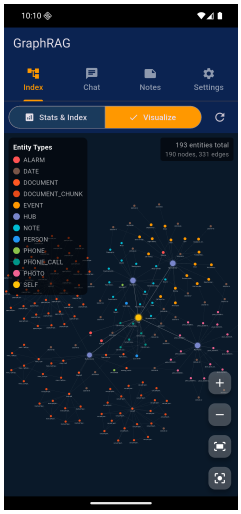
One of the most significant advantages of the RUVA architecture is user-controlled knowledge deletion. In vector-only systems, deleting a concept is mathematically imprecise, leaving behind probabilistic ghosts. In RUVA, since every memory is a discrete node in a relational database, deletion is exact. When a user chooses to forget a specific memory (e.g., "The project meeting on Friday"), RUVA executes a standard SQL `DELETE CASCADE` transaction targeting the reified information object. The node, its properties, and associated vector indices are permanently excised from the memory, granting users deterministic control over their digital past.

3 Demonstration Scenarios

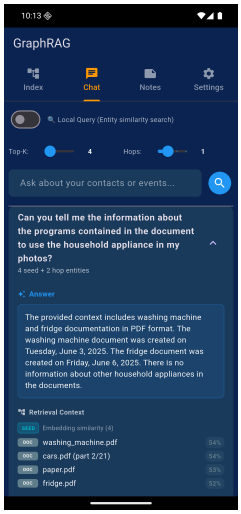
We demonstrate the "Glass Box" RUVA architecture with three use cases: (i) ingestion of a multimodal memory; (ii) complex reasoning, and (iii) memory deletion.

The demonstration begins with **Multimodal Ingestion (Scenario 1)**, illustrating how the system reacts when multimodal data enters the system. Let us introduce an example. The user syncs a calendar containing a "Weekend Trip" and uploads a photo of a train ticket. The *Multimodal Ingestion Workflow* triggers: the local VLM converts the ticket pixels into a structured node (`:Receipt :amount "95 EUR"`), while the system identifies the temporal overlap with the calendar event, forging a relationship edge.

Next, we demonstrate the **Complex Reasoning (Scenario 2)**. Again, let us show it through an example. The user asks: *"How much have I spent on the trip so far?"* Answering this requires bridging the concept "The Trip" from the calendar with granular data "95 EUR" from the image. The GraphRAG runtime executes a multi-hop traversal, locating

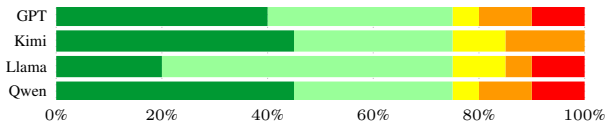


(a) Extracted PKG in RUVA

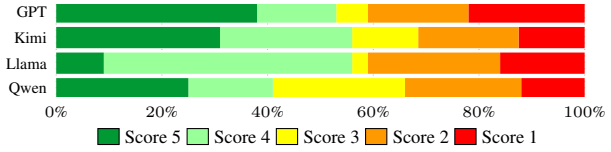


(b) Answer to multimodal query

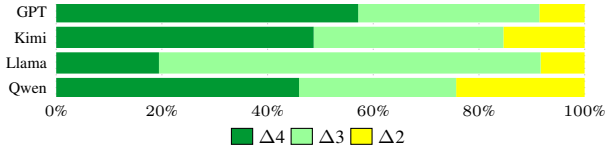
Figure 2: RUVA App Screenshots. Figure 2a shows the inferred PKG, while in Figure 2b RUVA answers leveraging multimodal data.



(a) Scenario 1: Multimodal Ingestion



(b) Scenario 2: Complex Reasoning



(c) Scenario 3: Deterministic Deletion (Delta Scores)

Figure 3: Judicial scores for **Scenario 1** and **2**, and Δ scores for **3**.

the Trip node and following the temporal edge to the Receipt node, answering *"You have spent 95 EUR for the ticket."*

Finally, we showcase the **Deterministic Deletion (Scenario 3)**, addressing the "Right to be Forgotten." In this example, the user decides that a financial piece of information (i.e., the receipt in Scenario 2) is sensitive and should not be retained. By selecting the *Receipt* in the gallery app and removing it, the system executes a deterministic `DELETE` cascade transaction. The node and all associated vector indices are excised. To verify this, the user asks the same question again, and the system responds: *"I couldn't find relevant information to answer your question."* This scenario demonstrates true data sovereignty: the user acts as the editor of their own memory, enforcing privacy with mathematical certainty in a way that opaque vector-only systems cannot guarantee.

4 Evaluation

This section assesses the RUVA's *operational feasibility* of the architecture and the *semantic reasoning quality*.

Performance. We deployed the system on a constrained environment, a Google Pixel 8 Pro. RUVA periodically scans the device's local data and compares it with existing nodes. Despite the constrained environment, RUVA ensures interactive latencies, with an average ingestion time of *2.4s* (including VLM processing and triple extraction) and a graph retrieval latency of only *38ms* for single-hop queries.

Model Accuracy. To comprehensively test RUVA, we prepared a benchmark comprising *71* multi-source objects (spanning calendar events, images, notes, textual documents, calls, alarms, and contacts) and 52 triplets containing questions, desired answers, and RUVA's answers. The benchmark specifically probes three distinct scenarios: *Multimodal Ingestion* (Figure 3a), *Complex Reasoning* (Figure 3b), and *Deterministic Deletion* (Figure 3c). The benchmark consists of **20 ingestion checks** and **32 complex reasoning** questions. To assess *Deterministic Deletion*, each question is evaluated with ($score_0$) and without ($score_1$) its corresponding information objects, measuring $\Delta = score_0 - score_1$. We considered answers whose judges' score is positive (3-5).

Experiments. We employ a panel of four state-of-the-art LLMs: Llama3.3-70B [Grattafiori et al., 2024], Qwen3-32B [Yang et al., 2025], GPT [Agarwal et al., 2025], and Kimi K2 [Bai et al., 2025] Instruct. Following the Prometheus [Kim *et al.*, 2024] strategy, judges evaluate benchmark triplets via Chain-of-Thought reasoning. To mitigate inter-model bias [Lee *et al.*, 2025], we map the resulting 5-point scores to three categories: *Positive* (4–5), *Neutral* (3), and *Negative* (1–2). RUVA demonstrates robust reasoning, with **61%** of responses rated as **Positive**, increasing to **71%** when including **Neutral** responses. (Figure 3a, Figure 3b). Reliability analysis shows strong cross-model agreement, with a Spearman's Rank Correlation (ρ) [Spearman, 1904] of **0.82** and a Krippendorff's Alpha [Krippendorff, 2011] of **0.81**, indicating consistency in ordinal evaluations. The **83% Percentage Agreement** [McHugh, 2012] and **Cohen's** κ [Cohen, 1968] of **0.81** confirm this finding. Notably, GPT and Kimi K2 achieve up to **92% agreement** ($\kappa = 0.92$), highlighting highly aligned assessments.

5 Conclusion

We introduce RUVA, redefining Personal AI from an opaque service into a transparent framework. By adopting a neuro-symbolic GraphRAG paradigm, we addressed the "Black Box" limitations that plague current retrieval systems. Achieving interactive latencies and high semantic accuracy on edge devices, RUVA proves the cloud is no longer a prerequisite for intelligence, ensuring data sovereignty. Its graph topology enables deterministic deletion, operationalizing the "Right to be Forgotten". By allowing users to inspect, correct, and erase memories, RUVA restores human agency.

Acknowledgments

The authors acknowledge partial support of OVS Fashion Retail Reloaded, Lutech Digitale 4.0, Servizi Locali 2.0, Konecta New KOCS ERA, P+ARTS, PNRR-MAD-2022-12376656, REACH-XY and CALLIOPE. Furthermore, this work has been carried out while Giovanni Servedio was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Politecnico di Bari.

References

[Agarwal et al., 2025] Agarwal et al. gpt-oss-120b & gpt-oss-20b model card. https://arxiv.org/abs/2508.10925, 2025. Accessed: 2026-06-10.

[Aghilar *et al.*, 2023] Potito Aghilar, Vito Walter Anelli, Michelantonio Trizio, and Tommaso Di Noia. Scalable cloud-native pipeline for efficient 3d model reconstruction from monocular smartphone images. In Bedir Tekinerdogan, Catia Trubiani, Chouki Tibermacine, Patrizia Scandurra, and Carlos E. Cuesta, editors, *Software Architecture - 17th European Conference, ECSA 2023, Istanbul, Turkey, September 18-22, 2023, Proceedings*, Lecture Notes in Computer Science, pages 266–282. Springer, 2023.

[Aghilar *et al.*, 2025] Potito Aghilar, Vito Walter Anelli, Michelantonio Trizio, Eugenio Di Sciascio, and Tommaso Di Noia. Training-free, identity-preserving image editing for fashion pose alignment and normalization. *Expert Syst. Appl.*, 293:128579, 2025.

[Anelli *et al.*, 2017] Vito Walter Anelli, Tommaso Di Noia, Pasquale Lops, and Eugenio Di Sciascio. Feature factorization for top-n recommendation: From item rating to features relevance. In Yong Zheng, Weike Pan, Shaghayegh (Sherry) Sahebi, and Ignacio Fernández, editors, *Proceedings of the 1st Workshop on Intelligent Recommender Systems by Knowledge Transfer & Learning co-located with ACM Conference on Recommender Systems (RecSys 2017), Como, Italy, August 27, 2017*, CEUR Workshop Proceedings, pages 16–21. CEUR-WS.org, 2017.

[Bai et al., 2025] Bai et al. Kimi k2: Open agentic intelligence. https://arxiv.org/abs/2507.20534, 2025. Accessed: 2026-06-10.

[Balog and Kenter, 2019] Krisztian Balog and Tom Kenter. Personal knowledge graphs: A research agenda. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2019, Santa Clara, CA, USA, October 2-5, 2019*, pages 217–220. ACM, 2019.

[Cohen, 1968] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

[Di Palma *et al.*, 2025] Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. Llamas have feelings too: Unveiling sentiment and emotion representations in llama models through probing. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 6124–6142. Association for Computational Linguistics, 2025.

[Edge *et al.*, 2024] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130, 2024.

[Grattafiori et al., 2024] Grattafiori et al. The llama 3 herd of models. https://arxiv.org/abs/2407.21783, 2024. Accessed: 2026-06-10.

[Kamath et al., 2025] Kamath et al. Gemma 3 technical report. https://arxiv.org/abs/2503.19786, 2025. Accessed: 2026-06-10.

[Kautz, 2022] Henry A. Kautz. The third AI summer: AAAI robert s. engelmore memorial lecture. *AI Mag.*, 43(1):93–104, 2022.

[Kim *et al.*, 2024] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models, 2024.

[Krippendorff, 2011] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011. Accessed: 2026-06-10.

[Lee *et al.*, 2025] Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15771–15798, Suzhou, China, November 2025.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

[Liang *et al.*, 2025] Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. The widespread adoption of large language model-assisted writing across society. *Patterns*, 6(12):101366, 2025.

[Luo *et al.*, 2025] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[McHugh, 2012] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012.

[Servedio *et al.*, 2025] Giovanni Servedio, Alessandro De Bellis, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. Are the hidden states hiding something? testing the limits of factuality-encoding capabil-

ities in llms. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 6089–6104. Association for Computational Linguistics, 2025.

[Skjæveland *et al.*, 2024] Martin G. Skjæveland, Krisztian Balog, Nolwenn Bernard, Weronika Lajewska, and Trond Linjordet. An ecosystem for personal knowledge graphs: A survey and research roadmap. *AI Open*, 5:55–69, 2024.

[Spearman, 1904] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, January 1904.

[Traag *et al.*, 2019] Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), March 2019.

[Wang *et al.*, 2023] Shoujin Wang, Yan Wang, Fikret Sivrikaya, Sahin Albayrak, and Vito Walter Anelli. Data science for next-generation recommender systems. *Int. J. Data Sci. Anal.*, 16(2):135–145, 2023.

[Wang *et al.*, 2025] Shang Wang, Tianqing Zhu, Dayong Ye, and Wanlei Zhou. When machine unlearning meets retrieval-augmented generation (rag): Keep secret or forget knowledge? *IEEE Transactions on Dependable and Secure Computing*, page 1–16, 2025.

[Yang et al., 2025] Yang et al. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. Accessed: 2026-06-10.