# H.E.M.A.: A Multimodal Cognitive Architecture and Hybrid Episodic Memory Framework for Human State-Aware Adaptive Behavior

Miguel Escudero-Jiménez
*Universidad Pablo de Olavide*
*Seville, Spain*
mescjim@upo.es

Eloy Retamino
*Universidad Pablo de Olavide*
*Seville, Spain*
ercarrion44@gmail.com

Randy Gomez
*Honda Research Institute, Japan*
*8-1 Honcho, Wakō-shi, Saitama 351-0188*
r.gomez@jp.honda-ri.com

Luis Merino
*Universidad Pablo de Olavide*
*Seville, Spain*
lmercab@upo.es

***Abstract*—Social robotics requires autonomous agents to maintain a persistent, multimodal understanding of a user's conversational, emotional, and physical state over long-term deployments. While state-of-the-art vision-language models (VLMs) offer impressive zero-shot understanding, they operate as stateless functions, causing context window degradation and an inability to track multi-hop relational history. This extended abstract presents H.E.M.A. (Hybrid Episodic Multimodal Architecture), an end-to-end cognitive framework natively built on ROS 2 for long-term user alignment. The architecture fuses high-frequency biometric, gaze, 3D body pose, and audio streams with asynchronous VLMs to extract rich tracking features into an ontology-validated Dynamic Scene Graph. This human state model integrates into a hybrid episodic memory system combining dense vector retrieval with a bounded Personalized PageRank graph walk. Simulation-based evaluations using a controlled interaction corpus and an LLM-as-a-Judge protocol show that our relational approach maintains real-time conversational latencies (averaging $95.85\,\mathrm{ms}$ for search) while drastically enhancing identity traceability ($+0.448$) and uncertainty management ($+0.460$) compared to flat vector baselines. Finally, we illustrate how this cognitive framework facilitates closed-loop behavioral adaptation, regulating physical gaze transitions through real-time multimodal speaker tracking and providing conversational pacing modulation via integrated, latency-based backchanneling policies.**

***Index Terms*—Social Robotics, Human-Robot Interaction, Multimodal Data Fusion, Hybrid Episodic Memory, Human-Adaptive Behavior, ROS 2.**

## I. Introduction and Objectives

### A. Background and Problem Statement

Tracking and understanding human states in real time is critical for natural Human-Robot Interaction (HRI), encompassing non-verbal multi-modal cues—such as eye gaze, 3D body orientation, facial expressions, and spatial presence—alongside spoken intent [1]. Translating these fleeting inputs into stable interactions requires a clear division between immediate situational context and consolidated personal experiences, mirroring the neurocognitive distinction between semantic and episodic memory models [2].

Although modern Large Language Models (LLMs) and Vision-Language Models (VLMs) provide robust zero-shot reasoning, they function as stateless entities [3], [4]. They lack native mechanisms to retain interaction history across sequential turns or sessions; yet, preserving this relational context across extended deployments is essential for grounding continuous, long-term human-adaptive behavioral adaptation strategies [5], [6]. Standard memory solutions rely on flat vector databases (VectorRAG) to retrieve past dialogue transcripts via cosine similarity [7], [8]. In human state-aware robotics, this baseline destroys temporal hierarchies, drops cross-turn context, and fails to answer multi-hop relational questions across distinct encounters [9], [10]. Conversely, unconstrained knowledge graph searches introduce computational latencies that violate conversational turn-taking boundaries, where system delays exceeding $2.5\,\mathrm{s}$ critically degrade human engagement [11], [12].

### B. Objectives

To resolve these challenges under physical, resource-constrained environments, this work introduces H.E.M.A. (*Hybrid Episodic Multimodal Architecture*), a platform-agnostic cognitive framework developed under the ROS 2 middleware ecosystem [13] and configured to meet the operational requirements of the Haru (*Honda Research Institute*) social robot [14]. The core objectives of this study are:

1) To construct an expressive, real-time representation of human conversational and physical states by fusing high-frequency tracking features with slower visual-semantic abstractions.
2) To organize these multimodal states into a long-term hybrid episodic memory system that bridges dense vector indexing with a bounded graph diffusion algorithm.

3) To validate the computational feasibility and semantic accuracy of the retrieval loop under episodic stress through a controlled ablation simulation, ensuring compliance with real-time HRI latency constraints.

## II. Methodology

The cognitive architecture is structured into three layers to isolate processing frequencies and maintain real-time conversational loops. The system interaction topology is illustrated in Fig. 1.

### A. Multimodal Data Fusion and State Synthesis (SENSE)

The SENSE layer utilizes a dual-pipeline processing architecture:

- **Pipeline A (Base Perception):** Pulls data from an RGB-D camera and directional microphone arrays. Integrated within a unified face perception package (Uniface [15]), the visual sub-system executes face detection through YOLOv8-Face, eye-gaze tracking, and ArcFace 128-D biometric extraction [16], smoothing identities via a majority-vote cache. The kinematic component runs YOLO26 pose estimation [17], lifting 2D landmarks into 3D camera space via the synchronized depth channel. An acoustic module extracts voice features using an ECAPA-TDNN architecture via a shared audio processing library [18].
- **Pipeline B (Semantic Understanding):** Runs an asynchronous local serving engine [19] hosting a VLM (Qwen3-VL [20]). Input frames undergo an optimized *person gating* crop to isolate the active speaker and reduce environmental noise. The prompt orchestration module overlays known biometrics as visual text hints on the raw image canvas to guide attention without degrading the structured JSON outputs.

To resolve speaker ties in noisy spaces, H.E.M.A. correlates an eye-normalized Mouth Aspect Ratio (MAR) variance over a rolling window [21] with the rising edge of audio Voice Activity Detection (VAD) [22]. Confirmed cross-modal alignment flags the participant as corroborated.

These streams feed into a central context fusion layer that merges low-level coordinates with deep visual-semantic descriptors into an in-memory *Dynamic Scene Graph* implemented via NetworkX [23]. 2D boundaries are enriched with 3D spatial vectors by re-projecting coordinate arrays against camera intrinsics. To handle object changes smoothly, every node property details its tracking source, classification confidence, and timestamp within a uniform *observable property framework*. Instead of destructive deletions, lost tracks decay via a *staleness lifecycle with a configurable Time-To-Live (TTL)*, preserving identity continuity across temporary occlusions. Every entity and relationship is validated against a formal Web Ontology Language (OWL) schema [24] prior to context publication.

### B. Cognitive Control and Intention Management (INTERACTION)

The INTERACTION layer routes incoming queries through a hierarchical, four-tier classification cascade designed to protect the active context window:

1) **Priority Lane:** Intercepts safety-critical words using regular expressions to immediately halt ongoing physical actions or verbal synthesis.
2) **Fast-Path Heuristics:** Directly resolves standard greetings, introductions, or deterministic profile queries without model execution overhead, routing predictable queries immediately to their matching execution tracks.
3) **Spatial Upgrade:** Scans the live scene graph for highly confident human-object relationships (`isHolding`, `isPointingAt`). If a match is found, the turn bypasses historical databases and upgrades directly to an environmental query.
4) **Model-Based Fallback:** Delegates ambiguous queries to a local language model (Qwen2.5-3B [25]) to classify the underlying intent. Requests with classification confidence scores below $0.55$ divert automatically to a designated unknown intent path.

To enforce state-dependent behavioral adaptation, the orchestrator dynamically channels the classified intents into three distinct execution routes:

- **Reactive Path:** Intercepts safety-critical words from the Priority Lane and basic conversational responses from Fast-Path Heuristics, resolving these turns instantly while completely bypassing episodic memory lookups and heavy visual processing channels.
- **Current Scene Path:** Triggered by Spatial Upgrades or explicit environmental queries identified by deterministic constraints or scene subtypes from the Model-Based Fallback. A visual anchor gate evaluates the live dynamic scene graph to resolve the turn via active relational edges, executing an asynchronous visual question answering (VQA) task only if the active graph data is insufficient or obsolete.
- **Historical Memory Path:** Processes queries regarding past events, persistent facts, or evolving preferences, routed either directly via deterministic fast-path heuristics or categorized under memory subtypes by the Model-Based Fallback. The system adapts by utilizing linear semantic search for simple chronological lookups and scaling to bounded Personalized PageRank (PPR) graph diffusion [26], [27] for complex multi-hop relational questions.

This persistent tracking directly drives two core execution axes for the target Haru robot's human-adaptive behavior. First, **Social Gaze Adaptation** utilizes a specialized head orientation control subsystem [28] to guide physical tracking across four modes (robot speaking, active speaker tracking, silence stabilization, and geometric fallback) based on cross-modal metrics and user proximity. Second, **Configurable Conversational Pacing** mitigates internal processing delays triggered by complex graph walks or VQA loops by deploying

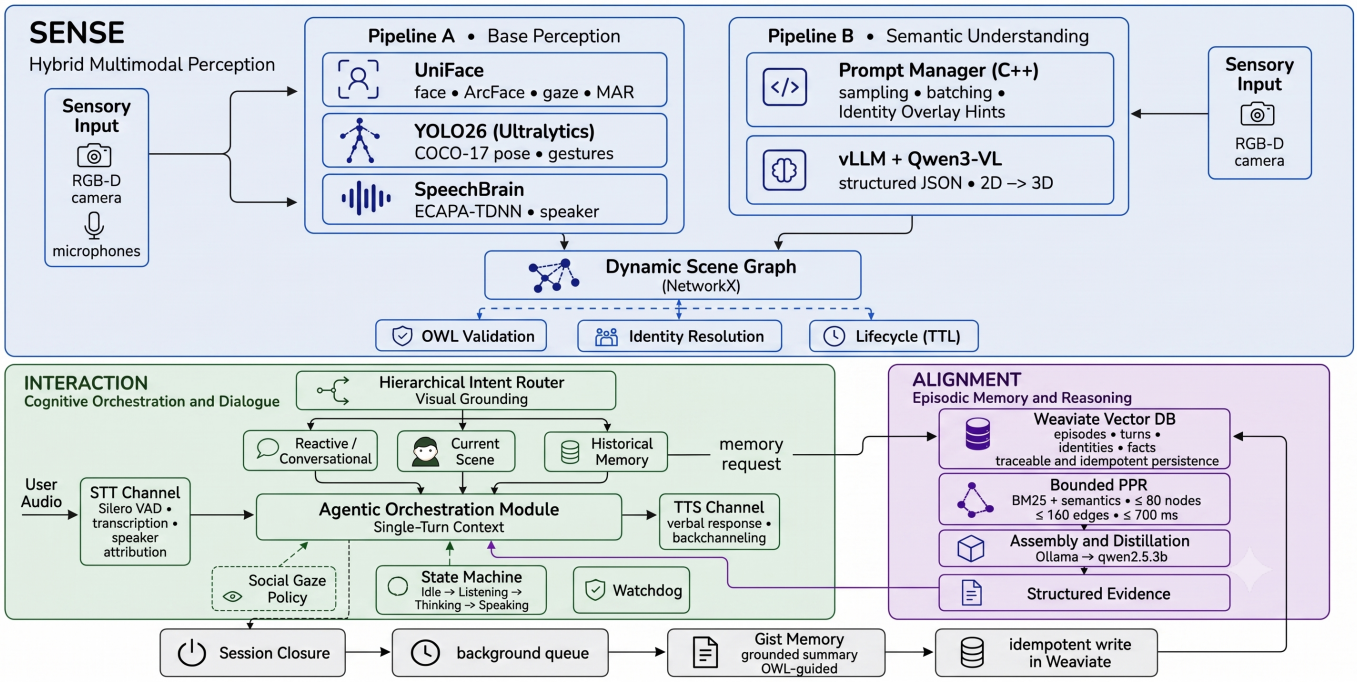


Fig. 1. The end-to-end cognitive pipeline of the H.E.M.A. framework, illustrating the layered isolation between high-frequency perception (SENSE), intention orchestration (INTERACTION), and vector-backed relational graph diffusion (ALIGNMENT) over a distributed robotic graph.

short continuity phrases and filler words to mask computational latency and preserve dialogue fluidity.

### C. Long-Term Memory and Relational Walk (ALIGNMENT)

The ALIGNMENT layer provides historical persistence through a local Weaviate instance [29], aligning multi-modal observations over extended timescales [30]. At session closure, the system filters out empty encounters and routes longer exchanges to a gist extraction pipeline [31] to produce interaction event anchors. Concurrently, an episodic visual memory subsystem archives keyframes vectorized via cross-modal vision-language encoders (CLIP/SigLIP) along with a normalized saliency score, enabling visual-to-visual similarity lookups.

Profile modifications and evolving user preferences update a durable fact index. The system standardizes keywords into canonical keys and checks vector distances to perform semantic no-op deduplication. When a user explicitly updates a preference, a targeted conflict resolution routine appends a retraction badge and audit metadata to the outdated information rather than deleting it, preserving historical accuracy.

For complex, multi-hop lookups, H.E.M.A. runs a **Bounded Personalized PageRank (Bounded PPR)** engine. The system extracts seed keywords from the query to build a local, multi-linked subgraph in RAM from historical database associations. PageRank activation propagates across this subgraph with a damping factor of $\alpha = 0.85$. To adhere to conversational constraints, the graph walk is bounded at $N_{\max} = 80$ nodes or $E_{\max} = 160$ edges and monitored by a 700 ms watchdog timer, falling back to flat vector search upon timeout.

## III. Results and Discussion

The cognitive stack was evaluated via an automated diagnostic ablation study on an interaction corpus, isolating the memory loop from physical hardware fluctuations. The experiment compared a *Forced Policy* (locking retrieval to either linear vector or Bounded PPR paths) against a *Free Policy* (allowing autonomous router selection). Metrics were audited post-hoc via an independent evaluation suite [32] with a high-parameter foundation model [33] serving as a blind judge.

### A. Quantitative Latency and Prompt Footprints

The execution metrics show that the relational graph layer introduces no latency penalties. Under the Bounded PPR path, net search and graph diffusion latency averaged $95.85 \pm 53.60$ ms (median: 88.40 ms), statistically equivalent to the flat linear semantic search baseline of $99.07 \pm 50.19$ ms ($p = 0.2165$). Hard topological constraints effectively cut off graph explosions, capping maximum search times at 297.73 ms. Total interactive turn latency averaged 2169.32 ms, safely remaining below the 2.5 s real-time boundary [11].

Furthermore, the graph walk optimized context assembly, reducing input prompt sizes during multi-hop queries to $4477.22 \pm 1446.27$ tokens—saving an average of 292.41 tokens per turn compared to the linear baseline ($p < 0.001$). Backend interface and network states monitored via our orchestration instrumentation confirm that localized bounding subgraphs maintain runtime latency stability across extended sessions.

### B. Qualitative Alignment and Behavioral Adaptation

Under deep chronological stress (dense interaction sessions with shifting user preferences and roles over time), the flat vector space suffered from semantic overlap, causing context bleed. Table I summarizes the qualitative metrics provided by the independent judge.

TABLE I
MODEL-AS-A-JUDGE PERFORMANCE COMPARISON UNDER DEEP CHRONOLOGICAL STRESS (NORMALIZED RANGE $[0, 1]$).

| **Metric Dimension** | **Linear Semantic** | | **H.E.M.A. (PPR)** | | $p$**-value** |
|---|---|---|---|---|---|
| Contextual Recall | 0.375<br>0.454 | $\pm$ | **0.780**<br>**0.389** | $\pm$ | $< 0.001$ |
| Traceability | 0.304<br>0.421 | $\pm$ | **0.752**<br>**0.373** | $\pm$ | $< 0.001$ |
| Uncertainty Handling | 0.275<br>0.405 | $\pm$ | **0.735**<br>**0.382** | $\pm$ | $< 0.001$ |
| Historical Fidelity | **0.802**<br>**0.315** | $\pm$ | 0.769<br>0.330 | $\pm$ | 0.0727 |

The graph-based path significantly outperformed the linear baseline, achieving absolute increases of over $+0.40$ points in *Contextual Recall*, *Traceability*, and *Uncertainty Handling*. These structural memory advantages directly ground the framework's broader deployment capabilities, smoothly feeding the social gaze adaptation, configurable conversational pacing, and cross-modal grounding mechanisms detailed in Sec. II.

Additionally, high uncertainty scores empower the system to handle knowledge gaps safely; instead of inventing data, the agent explicitly states the limits of its memory (e.g., acknowledging a past object but omitting an untracked donor). However, evaluating the *Free Policy* revealed a major routing bottleneck: the intent router misclassified $84.4\%$ of multi-hop queries, sending them to the linear semantic path by default and highlighting a key target for future optimization.

## IV. CONCLUSION

### A. Generalizations

The simulation-based evaluation of H.E.M.A. demonstrates that hybrid episodic memory frameworks can successfully structure and ground high-dimensional, multimodal human states without losing vital temporal or relational details. Fusing high-frequency sensory metrics into an OWL-validated Dynamic Scene Graph provides a stable context representation that seamlessly supports predictable, closed-loop behavioral adaptations like adaptive physical gaze tracking and latency-mitigating conversational backchanneling.

### B. Limitations and Future Work

The primary operational limitation identified in this study is the high misclassification rate ($84.4\%$) exhibited by the intent classifier under unconstrained conditions, which frequently routes complex multi-hop queries away from the relational graph path. Future work will focus on replacing this text-only router with a graph-guided routing mechanism that injects active ontology snippets directly into the local model's prompt context. Finally, full-scale population field trials will be conducted on the physical Haru platform to evaluate long-term user bonding, usability metrics, and perceived cognitive load during unconstrained interactions.

## ACKNOWLEDGMENT

This work was partially supported by the projects AI-FUSE (SAIA202500X163851SV0) and COBUILD (PID2024-161069OB-C31), funded by the Spanish Research Agency and the Ministry of Science and the European Union (MCIN /AEI/10.13039/501100011033) and by ERDF "A way of making Europe"

## REFERENCES

[1] Y. Wang, Y. Xu, A. Nikolova, Y. Wang, J. Wang, C. Wang, and X. Tong, "How do we research human-robot interaction in the age of large language models? a systematic review," *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '26)*, 2026.

[2] E. Tulving, "Episodic and semantic memory," in *Organization of Memory*, E. Tulving and W. Donaldson, Eds. New York, NY, USA: Academic Press, 1972, pp. 381–403.

[3] J. Atuhurra, "Leveraging large language models in human-robot interaction: A critical analysis of potential and pitfalls," 2024.

[4] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman, B. Ichter, D. Driess, J. Wu, C. Lu, and M. Schwager, "Foundation models in robotics: Applications, challenges, and the future," *The International Journal of Robotics Research*, vol. 44, no. 1, pp. 3–27, 2024.

[5] S. Paplu, R. F. Navarro, and K. Berns, "Harnessing long-term memory for personalized human-robot interactions," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, 2022, pp. 377–382.

[6] Y. Wu, S. Liang, C. Zhang, Y. Wang, Y. Zhang, H. Guo, R. Tang, and Y. Liu, "From human memory to ai memory: A survey on memory mechanisms in the era of llms," 2025.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," 2021.

[8] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024.

[9] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," 2023.

[10] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," 2025.

[11] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021.

[12] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon, and S. C. Levinson, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009.

[13] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot Operating System 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, p. eabm6074, 2022.

[14] R. Gomez, D. Szapiro, S. Cooper, N. Bougria, G. Pérez, E. Nichols, J. Giménez-Figueroa, J. M. Perez-Moleron, M. Peavy, D. Serrano, and L. Merino, "Design of embodied mediator haru for remote cross cultural communication," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 5505–5511.

[15] J. Yakhyo, "UniFace: A Unified Framework for Face Detection, Recognition, and Gaze Estimation," https://github.com/yakhyo/uniface, 2024.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[17] G. Jocher and J. Qiu, "Ultralytics yolo26," 2026. [Online]. Available: https://github.com/ultralytics/ultralytics

[18] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[19] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Z. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with PagedAttention," in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (SOSP)*, 2023, pp. 611–626.

[20] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-vl technical report," 2025.

[21] T. Soukupová and J. Cech, "Real-time eye blink detection using facial landmarks," 2016.

[22] Silero Team, "Silero VAD: Pre-trained enterprise-grade voice activity detector," https://github.com/snakers4/silero-vad, 2021.

[23] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," *Python in Science Conference*, 2008.

[24] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein, "OWL web ontology language reference," W3C, W3C Recommendation, 2004.

[25] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115

[26] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking : Bringing order to the web," in *The Web Conference*, 1999. [Online]. Available: https://api.semanticscholar.org/CorpusID:1508503

[27] G. Jeh and J. Widom, "Scaling personalized web search," *Proceedings of the 12th International Conference on World Wide Web, WWW 2003*, 08 2002.

[28] Y. Fang, J. M. Pérez-Molerón, L. Merino, S.-L. Yeh, S. Nishina, and R. Gomez, "Enhancing social robot's direct gaze expression through vestibulo-ocular movements," *Advanced Robotics*, vol. 38, no. 19-20, pp. 1457–1469, 2024. [Online]. Available: https://doi.org/10.1080/01691864.2024.2398556

[29] Weaviate B.V., "Weaviate: AI-native vector database," https://weaviate.io/, 2024.

[30] L. Long, Y. He, W. Ye, Y. Pan, Y. Lin, H. Li, J. Zhao, and W. Li, "Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory," 2025.

[31] P. Zhou, H. Li, Z. Nie, J. Chen, Q. Gong, W. Zhang, and C. Yu, "Understand then memory: A cognitive gist-driven rag framework with global semantic diffusion," 2026.

[32] Confident AI, "DeepEval: The open-source LLM evaluation framework," https://deepeval.com/, 2025.

[33] Google AI for Developers, "Gemini Models: Gemini API documentation," https://ai.google.dev/gemini-api/docs/models, 2026.