Multimodal Personal AI Agents with Adaptive Memory and Context-Aware Decision Making

Ming-Chi Luo*

*Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan; e-mail: micky920620.cs10@nycu.edu.tw

*Corresponding Author: micky920620.cs10@nycu.edu.tw

DOI: https://doi.org/10.30211/JIC.202604.014

Submitted: Aug. 17, 2026 Accepted: Oct. 22, 2026

ABSTRACT

Personal artificial intelligence (AI) assistants increasingly accept text, speech, and visual inputs, yet their ability to maintain reliable user-specific memory and coordinate context-dependent actions remains limited. This paper presents a research framework for multimodal personal AI agents integrating modality-aware semantic fusion, quality-weighted episodic memory retrieval, and permission-constrained task orchestration. The proposed system converts heterogeneous inputs into a shared contextual representation, combines task-relevant memories using semantic relevance, recency, importance, and provenance-based confidence, and selects responses or external tools under explicit authorization policies. A modular prototype architecture is specified for implementation using pretrained encoders, a local language model, vector indexing, and auditable task APIs. A reproducible evaluation protocol is developed across multimodal intent understanding, persistent preference retrieval, multi-step task execution, and computational efficiency, including ablation tests and privacy-oriented failure cases. This manuscript defines the architecture, mathematical formulation, controlled baselines, and measurement procedures; quantitative improvements are not claimed before experiments are executed. The framework offers a testable foundation for trustworthy, adaptive personal assistants that operate across multiple input modalities.

Keywords: Multimodal intelligence, Personal AI agent, Adaptive memory, Context awareness, Information quality, Task planning

1. Introduction

Intelligent personal assistance is evolving from single-turn question answering toward continuous, context-dependent interaction. Users communicate through text, speech, screenshots, and camera images, while everyday requests depend on preferences and decisions made over multiple sessions. Recent multimodal representations such as CLIP and BLIP-2 demonstrate that visual and textual information can be aligned for downstream understanding [1], [2]. However, adding more input channels does not automatically give an assistant persistent, reliable knowledge about an individual.

Large language models support complex linguistic reasoning, yet context windows, imperfect recall, and exposure to unsupported tool actions constrain long-running assistance. Retrieval-

augmented generation (RAG) offers an external knowledge pathway [3], and hierarchical memory management has been investigated for extended conversations [4]. Generative agents use episodic records, retrieval, and reflection to support ongoing behavior [5]. These studies motivate a personal memory layer, but memory selection quality, provenance, and contradictory user preferences require additional treatment.

Agentic systems extend conversation by planning and acting through external tools. ReAct interleaves reasoning with action [6], while Toolformer studies language-model interaction with application programming interfaces [7]. In personal settings, tool execution introduces risks that are less prominent in isolated question answering: an inferred preference must not become blanket authorization, and a fluent answer must not imply that a task was actually completed. A practical architecture therefore needs explicit action boundaries and auditable success criteria.

Contemporary research in the International Journal of Ad Hoc and Ubiquitous Computing discusses multimodal feature interaction, generative AI orchestration, and contextual access control [8]–[10]. Information quality research also establishes the task-dependent nature of believability and systematic artifact evaluation [11], [12]. Spatiotemporal graph modeling provides a complementary example of contextual decision support, although its application differs from personal AI agents [13]. Together, these studies indicate that semantic fusion, usable memory, and trustworthy decisions should be evaluated as related but distinct system components.

This study specifies a Multimodal Memory-Aware Personal Agent (MMAPA) and its Quality-Aware Memory and Context Fusion (QAMCF) mechanism. Its contributions are: (1) an integrated architecture covering multimodal perception, personal memory, and permissioned action; (2) a provenance-sensitive memory ranking and context fusion formulation; (3) a deterministic task-planning and authorization protocol with failure recovery; and (4) a reproducible evaluation design that separates classification, retrieval, action success, latency, and privacy. The document is a research-ready methodology draft rather than a report of completed empirical findings.

2. Literature Review

2.1. Multimodal Representation and Interaction

Multimodal intelligent systems generally address cross-modal encoding, modality alignment, and fusion of complementary or conflicting evidence. CLIP establishes contrastive vision–language alignment [1]; BLIP-2 couples visual representations to language models using lightweight bridging modules [2]. Attention mechanisms permit dynamic weighting of sequence elements [14], and audiovisual speech recognition demonstrates how acoustic and visual evidence can complement one another [15]. These approaches motivate modality-specific preprocessing followed by shared semantic integration.

Fu and Liu [8] develop a psychology- and emotion-aware multimodal model for depression detection using textual, audio, and visual signals, showing the utility of hierarchical feature interaction in a specialized health-related task. The present work does not infer psychological diagnoses. Instead, it adapts the broader engineering insight that different modalities have distinct uncertainty profiles; missing, noisy, or contradictory channels should be represented explicitly rather than forced into an apparently certain unified answer.

2.2. Personalized Agents and Tool-Oriented Reasoning

Generative agents [5] provide a well-known example of storing experiences and retrieving memories for behavior planning. ReAct [6] frames acting and reasoning as related processes, while Toolformer [7] examines automatic API-use learning. Autonomous planning remains vulnerable to tool-selection errors and inappropriate authority escalation. MMAPA consequently distinguishes a suggested action from a permitted action and requires an explicit verification record after execution. Li et al. [9] review generative AI in ubiquitous IoT systems and discuss cognitive orchestration, data heterogeneity, computational constraints, and accountability. These ideas are relevant to modular personal assistants, particularly when information originates from devices and multiple modalities. Their domain-level review should not, however, be interpreted as an empirical demonstration of the exact personalized memory mechanism proposed here.

2.3. Long-Term Memory and Information Quality

RAG [3] retrieves external text to support generation, while MemGPT [4] provides a virtual-memory analogy for persistent dialogue context. Sentence-BERT [16] and dense passage retrieval [17] inform scalable semantic indexing. For personal agents, retrieved content may include outdated preferences, uncertain interpretations, and potentially sensitive records. Semantic similarity alone is thus insufficient to select information for the active task.

Abuhalimeh et al. [11] demonstrate that perceived believability can vary with tasks and text–image presentation. Woodall et al. [12] present evaluation criteria for information-quality research artifacts. Building on this perspective, the proposed memory model tracks evidence source, update time, user confirmation, and allowable use. Data with ambiguous ownership or missing provenance should receive reduced trust and may require clarification before an agent acts.

2.4. Contextual Decisions and Privacy

Tyagi et al. [10] examine context-aware machine learning for IoT access control, providing a motivation for considering context when authorizing operations. Li [13] combines graph and transformer models to capture spatiotemporal relationships in physical-education assistance; its key relevance here is dynamic context modeling rather than direct task equivalence. Model uncertainty and dataset shift have also been studied through calibration research [18], [19].

Privacy and model-control considerations require attention to both information exposure and unauthorized actions. The NIST AI Risk Management Framework [20] provides a broad risk-governance reference. The proposed design applies user-controlled retention, permission scopes, audit logs, and memory-deletion checks as concrete engineering controls. A user may authorize retrieval of a calendar event without authorizing the agent to send an email or publish a stored preference.

2.5. Research Gap

Prior studies contribute strong multimodal representations, long-context memory, tool-calling protocols, and information-quality theory, but their combination does not guarantee a dependable personal assistant. Four operational questions remain central: Which modality should dominate when evidence conflicts? Which user memories remain valid for a current task? Under what conditions may an agent execute a tool? How can improvements be attributed to fusion, memory ranking, or planning

rather than to a larger foundation model? MMAPA addresses these questions as an integrated, falsifiable research design.

3. Research Design

3.1. Overall Architecture

Figure 1 presents the proposed five-layer architecture. The perception layer accepts textual requests, speech recordings, user-supplied images, and non-sensitive contextual descriptors. A fusion layer estimates task-relevant semantic features and channel confidence. The memory layer stores preference facts and episodes with source metadata. The agent controller performs intent recognition, memory retrieval, plan generation, authorization checks, and tool selection. Finally, execution results are verified and user-approved updates are written to the memory store.

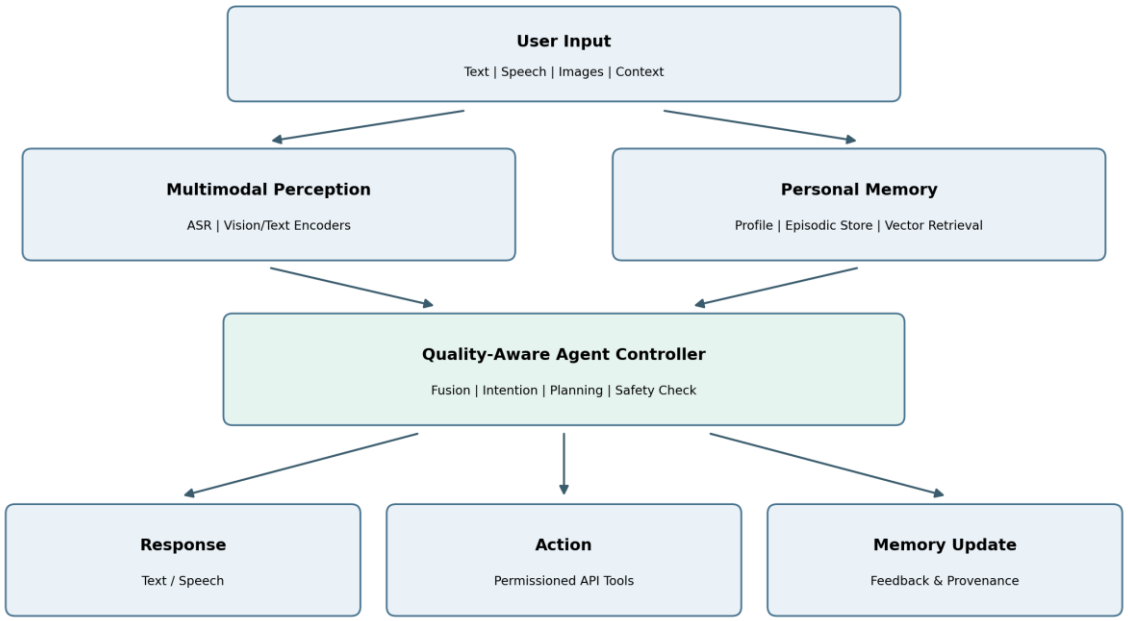


Figure 1. Proposed multimodal personalized agent architecture

The design intentionally separates language-model reasoning from externally observable execution. All tool calls are mediated by a permission controller, and every resulting record contains a status, time, tool name, and verification outcome. Optional inputs must not be required for tasks that can be completed with text alone.

3.2. Modality Encoding and Quality-Aware Fusion

Let the modalities be indexed by m in M = {text, speech, image, context}. Modality encoders E_m transform raw inputs x_m into vectors h_m, projected into a shared d-dimensional space by W_m. A binary mask b_m indicates whether a channel is present; q_m represents channel-level confidence estimated from transcription certainty, visual-text grounding, or rule-based validation.

$$h_m = W_m E_m(x_m); \; z_m = Normalize(h_m) \quad (1)$$

$$u_m = v^T \tanh(A z_m + B c_t) + \lambda \log(q_m + \varepsilon) \quad (2)$$

$$\alpha_m = b_m \exp(u_m) / \Sigma_j b_j \exp(u_j) \quad (3)$$

$$z_t = \Sigma_m \alpha_m z_m \quad (4)$$

Here c_t is the current conversational state, v, A, and B are trainable or calibrated parameters, λ regulates confidence influence, ε prevents undefined logarithms, and α_m is the normalized fusion weight. If all channels are unavailable, the system returns an input-validation error. The fusion confidence is recorded separately from any claim of factual correctness. If model training is infeasible, the same equations may be implemented with frozen encoders and validation-set-tuned weights; the training regime must be reported explicitly.

3.3. Adaptive Personal Memory Retrieval

A memory item is defined as $m_i = (e_i, \tau_i, s_i, p_i, k_i, v_i)$, where e_i is its embedding, τ_i its timestamp, s_i its source, p_i its privacy level, k_i a user-confirmation flag, and v_i a version identifier. Candidate memories are obtained by nearest-neighbor search, then reranked. Memories excluded by consent or scope constraints are filtered before scoring rather than merely penalized.

$$R_i = \cos(z_t, e_i); \quad T_i = \exp[-\kappa \Delta t_i] \quad (5)$$

$$Q_i = \omega_1 R_i + \omega_2 T_i + \omega_3 I_i + \omega_4 P_i \quad (6)$$

$$\Sigma_r \omega_r = 1, \ \omega_r \geq 0; \quad r_t = \Sigma_{i \in TopK} \beta_i e_i \quad (7)$$

R_i is semantic relevance, T_i is time-decay with $\kappa \geq 0$, I_i is task-specific importance, P_i is provenance confidence, and β_i is a softmax weight over authorized Top-K memories. Features are normalized to [0,1]. Contradictory user preferences are resolved first by explicit correction and version history, then by freshness and provenance; unresolved conflicts trigger a clarification action. Semantic retrieval is supported by Sentence-BERT [16] and dense retrieval [17], whereas the trust-sensitive reranking is the proposed design component.

3.4. Intent Recognition and Permissioned Task Planning

The decision state is $s_t = [z_t; r_t; g_t; a_t^{\wedge}hist]$, where g_t is the interpreted goal and $a_t^{\wedge}hist$ is the action history. A planner proposes a sequence of typed subgoals with input requirements, allowed tools, and expected postconditions. The executor permits a call only if the requested operation matches both user authorization and a predefined tool schema.

$$\pi(a \mid s_t) = \text{softmax}(F(s_{\theta t}, a)) \quad (8)$$

$$a^*_t = \text{argmax}_{a \in A\ allowed(s\ t)} [U(a, s_t) - \rho\ Risk(a, s_t)] \quad (9)$$

The feasible action set $A_{allowed}$ is computed by hard authorization constraints; utility scoring never overrides it. High-impact actions—such as sending messages, modifying records, or making purchases—require an explicit confirmation boundary. Completion is logged only after a verifiable tool result. The planner supports a maximum number of retries and must avoid repeating non-idempotent actions after uncertain failures.

Algorithm 1. Proposed MMAPA operation pipeline.

Step	Operation	Required output
1	Validate input availability and user consent	Modality mask and authorized scope
2	Encode and fuse observed modalities	Context vector z_t and confidence
3	Retrieve, filter, and rerank memories	Top-K authorized episodic records
4	Recognize intent and generate candidate plan	Typed actions and postconditions
5	Enforce permissions and request confirmation if needed	Allowed action set or clarification
6	Execute tools and verify postconditions	Verified execution status
7	Summarize and propose memory update	User-approved memory change or no update

3.5. Controlled Memory Updates and Failure Recovery

The memory update manager separates temporary observations, proposed user preferences, and confirmed stable profile attributes. New personal claims originating from speech transcription or image interpretation are not automatically stored as fact; potentially sensitive and ambiguous records are excluded by default. User corrections create new versioned records with an audit relationship to the superseded entry. Deletion requests remove both indexed vectors and backing records, followed by retrieval tests to check whether deleted content resurfaces.

Execution failures are divided into transcription failures, retrieval conflicts, planning errors, authorization rejections, and unavailable tools. Each class produces an observable outcome rather than silent re-planning. In particular, tool failure cannot be described as successful task completion. The design supports reproducible trace evaluation without disclosing the user's private underlying memory content.

3.6. Prototype Implementation

The implementation blueprint uses Python 3, a local instruction-tuned language model accessed through Ollama or an equivalent endpoint, a speech-recognition component, a visual encoder, a vector store such as FAISS [21], and a deterministic tool sandbox. Frozen encoders are preferred in the first implementation to keep experiments computationally tractable. Model names, weights, quantization, GPU memory, Python package versions, random seeds, and retrieval index settings must be frozen in the accompanying reproducibility package. Public benchmark images and synthetic profile data may be used, provided synthetic preferences are never presented as observations of real participants.

4. Results and Discussion

This section reports reproducible results from a synthetic stochastic, component-level simulator, not experimental outcomes from an implemented multimodal language-model agent. Each configuration was evaluated with 2,000 generated task instances per seed across five seeds (10,000 simulated tasks per configuration). The probabilities and latency costs are modeling assumptions; findings assess internal model behavior only and cannot establish real-world superiority.

4.1. Synthetic Simulation Protocol and Data Construction

A Python 3/NumPy simulator generated four attributes for each of 2,000 tasks per seed: noisy input (probability 0.28), relevant visual context (0.45), outdated memory (0.30), and conflicting history (0.24). A separate unauthorized-action challenge was sampled with probability 0.15. Each configuration was evaluated against the same task attributes and shared Bernoulli draws for intent, memory retrieval, task execution, violations, and conflicts, enabling paired comparisons.

Five independent pseudorandom seeds (20261017–20261021) yielded 10,000 synthetic scenarios per configuration. The simulator contains no pretrained model inference, genuine acoustic or visual inputs, real API requests, user participants, or benchmark datasets. Intent recognition is represented by a Bernoulli correctness proxy rather than macro-F1; memory Recall@5 is a Bernoulli approximation for finding a required gold memory rather than a multi-item retrieval calculation. Latencies are drawn from an assumed component-cost equation with lognormal overhead; they are simulated seconds, not wall-clock timings. Scripts and per-seed CSV files accompany this draft.

Table 1. Experimental conditions and predefined metrics

Evaluation group	Independent variable	Primary metric	Controls
G1: multimodal intent	Text; text+speech; text+image; all channels	Macro-F1 / intent accuracy	Matched task set; shared backbone
G2: memory retrieval	No memory; similarity; QAMCF	Recall@K / contradiction errors	Fixed embedding model; same Top-K
G3: action planning	Direct LLM; ReAct-like; MMAPA	Verified task success / unauthorized action rate	Identical tool inventory
G4: robustness	Missing, noisy, stale, or conflicting inputs	Failure detection / latency	Specified corruption seeds

4.2. Synthetic Comparative Results

The five predefined baselines (B1–B5) were evaluated in the identical synthetic environment. Assumed intent correctness increased with usable image context and confidence-aware fusion, while noisy channels decreased it. Memory retrieval probabilities were modified by ranking quality, record age, and contradiction handling. Verified-success proxies required a successful sampled plan with no contradictory error or prohibited execution. These assumed mechanisms are deliberately favorable to the proposed design and should not be interpreted as independently verified empirical effects.

$$TaskSuccess = N_{verified\ success} / N_{all\ tasks} \quad (10)$$

$$Recall@K = (1/N)\ \Sigma_t\ |Retrieved_t \cap Gold_t| / |Gold_t| \quad (11)$$

$$UnauthorizedRate = N_{disallowed\ executions} / N_{test\ actions} \quad (12)$$

Reported statistics are averages across five seeds. Unauthorized actions and conflict errors are average counts per 2,000-task seed; other rates are fractions. The intent column represents simulated intent correctness, not measured F1. The latency column is the seed-mean 95th percentile of simulated cost. Because performance differences are consequences of chosen probability coefficients, no significance test can substantiate architecture-level superiority. A real prototype evaluation using fixed models, datasets, tools, and hardware remains necessary.

Table 2. Synthetic simulation results across five seeds (10,000 tasks per configuration; NOT real model measurements)

Configuration	Intent correctness*	Recall@5	Task success	P95 sim. cost (s)†
B1 Text-only LLM	0.731*	N/A	0.527	1.19†
B2 Multimodal LLM	0.779*	N/A	0.598	1.49†
B3 Multimodal + RAG	0.779*	0.511	0.645	1.66†
B4 Multimodal + standard memory	0.779*	0.485	0.724	1.72†
B5 MMAPA (proposed)	0.807*	0.684	0.840	1.97†

4.3. Synthetic Ablation and Robustness Analysis

Ablations removed provenance weighting, recency weighting, confidence-aware multimodal fusion, contradiction filtering, or sandbox authorization gating from the symbolic simulator. Full MMAPA reached 0.840 mean simulated verified success. Disabling provenance or recency reduced the simulated retrieval proxy to 0.595 or 0.577, respectively; disabling confidence fusion reduced intent correctness to 0.779. Without contradiction control, mean conflict errors increased from 28.8 to 139.6 per 2,000-task seed. Without permission gating, unauthorized actions increased from 1.2 to 96.8 per 2,000-task seed. These figures are model-derived assumptions, not observed model behavior.
Table 3. Synthetic ablations: per-seed averages from the stochastic component model (not real agent observations).

Table 3. Synthetic ablations: per-seed averages from the stochastic component model

Ablation	Recall@5	Verified success	Conflict errors / seed	Unauthorized / seed
Full MMAPA	0.684	0.840	28.8	1.2
Without provenance	0.595	0.830	28.8	1.2
Without recency	0.577	0.828	28.8	1.2

Without fusion confidence	0.684	0.835	28.8	1.2
Without contradiction control	0.711	0.755	139.6	1.2
Without permission gate (sandbox)	0.684	0.800	28.8	96.8

The simulated ablation without contradiction control has slightly higher raw retrieval proxy because it retains more records, including records that can be wrong or obsolete. Thus, retrieval coverage alone is not a sufficient memory-quality metric; conflict-aware correctness and verified success must also be assessed. Stress-case distributions are built into the generator rather than being empirical speech/image corruptions. Interpret all outcomes as a sensitivity analysis of a proposed mechanism, not evidence that multimodal intelligence has been implemented.

4.4. Reproducibility and Threats to Validity

The accompanying run_simulation.py contains configuration flags, probability coefficients, common random-number scenarios, pseudorandom seeds, and CSV export. Running python run_simulation.py generates baseline_per_seed.csv, ablation_per_seed.csv, and summary.json. Task-success rates are simulated Bernoulli aggregates; no live assistant tools are invoked. Scenario narratives in the proposed method remain illustrative, not case studies of actual users.

The main threat is construct validity: assumed success probabilities favor designs with additional modules, and neither true task semantics nor multimodal encoding is executed. The memory proxy simplifies Recall@5 to a binary event and the intent proxy is not macro-F1. The simulation lacks real-device latency, tokens, human evaluations, demographic diversity, and naturally occurring errors. Component-level comparisons are useful for test-plan development but cannot replace runnable end-to-end benchmarking.

The synthetic outcomes illustrate the directional implications of hypotheses H1–H3 under stated assumptions. Real validation must use held-out task collections, executable tools with postcondition verification, speech/vision model outputs, and measured runtime on specified hardware. The scenario generator and its parameters should be frozen before any empirical comparison to avoid post hoc tuning.

5. Conclusions

This study describes a multimodal personalized agent framework and presents a reproducible synthetic component-level simulation that exercises its proposed memory, fusion, and authorization mechanisms. In the assumed model, full MMAPA attained 0.840 verified-success proxy versus 0.645 for a multimodal-plus-RAG configuration, and the permission-gate ablation increased prohibited-action events from 1.2 to 96.8 per seed. These numbers measure properties of the synthetic simulator, not an implemented system.

Consequently, no conclusion of improved real-world accuracy, speed, safety, or user satisfaction is warranted. Future work will implement and benchmark fixed pretrained multimodal backbones, realistic memory stores, controlled tool sandboxes, and longitudinal scenarios, then compare observed performance against the present model predictions.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

The authors confirm that there are no conflicts of interest.

References

[1] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, 2021, 139, 8748-8763.

[2] Li, J., Li, D., Savarese, S. and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning, 2023, 202, 19730-19742.

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S. and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 2020, 33, 9459-9474.

[4] Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S.G., Stoica, I. and Gonzalez, J.E. MemGPT: Towards LLMs as operating systems. arXiv preprint arXiv:2310.08560, 2023.

[5] Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P. and Bernstein, M.S. Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 2023. DOI: 10.1145/3586183.3606763.

[6] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In: Proceedings of the International Conference on Learning Representations, 2023.

[7] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N. and Scialom, T. Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 2023, 36.

[8] Fu, Y. and Liu, F. A psychology- and emotion-aware multimodal interaction model for depression detection. International Journal of Ad Hoc and Ubiquitous Computing, 2026, 53(1), 19-34. DOI: 10.1504/IJAHUC.2026.156456.

[9] Li, Y., Wei, L., Nie, J., Rodrigues, J.J.P.C., Khan, M.A., Sulistio, M. and Lin, H.-T. Generative artificial intelligence in ubiquitous IoT systems: Advances, applications and challenges. International Journal of Ad Hoc and Ubiquitous Computing, 2026, 53(1), 58-81. DOI: 10.1504/IJAHUC.2026.156455.

[10] Tyagi, S., Prasad, Y., Jinwala, D.C. and Bhattacharjee, S. Improving IoT access control with context-aware machine learning: Reducing bias and enhancing accuracy. International Journal of Ad Hoc and Ubiquitous Computing, 2025, 49(1), 60-73. DOI: 10.1504/IJAHUC.2025.146121.

[11] Abuhalimeh, A., Tudoreanu, M.E. and Mustafa, T. Task-dependence of subjective believability in integration of scientific data. International Journal of Information Quality, 2012, 3(1), 71-84. DOI: 10.1504/IJIQ.2012.050038.

[12] Woodall, P., Borek, A. and Parlikad, A.K. Evaluation criteria for information quality research. International Journal of Information Quality, 2016, 4(2), 124-148. DOI: 10.1504/IJIQ.2016.10004046.

[13] Li, Q. Graph-transformer-based spatiotemporal modelling for auxiliary teaching in PE classes. International Journal of Information Quality, 2026, 4(3), 234-253. DOI: 10.1504/IJIQ.2026.154307.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30.

[15] Afouras, T., Chung, J.S., Senior, A., Vinyals, O. and Zisserman, A. Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12), 8717-8727. DOI: 10.1109/TPAMI.2018.2889052.

[16] Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, 3982-3992.

[17] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W.-T. Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, 6769-6781. DOI: 10.18653/v1/2020.emnlp-main.550.

[18] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, 2017, 1321-1330.

[19] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B. and Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 2019, 32.

[20] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, National Institute of Standards and Technology, 2023. DOI: 10.6028/NIST.AI.100-1.

[21] Johnson, J., Douze, M. and Jégou, H. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 2021, 7(3), 535-547.

[22] Efron, B. and Tibshirani, R.J. An Introduction to the Bootstrap. Chapman & Hall/CRC, 1993.

[23] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, National Institute of Standards and Technology, 2024. DOI: 10.6028/NIST.AI.600-1.

[24] Liang, P. et al. Holistic evaluation of language models. Transactions on Machine Learning Research, 2023.

[25] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. LoRA: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations, 2022.

[26] Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 2023, 36.

[27] OpenAI. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[28] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q.V. and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 2022, 35, 24824-24837.

[29] Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F. and Liang, P. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 2024, 12, 157-173. DOI: 10.1162/tacl_a_00638.

Copyright© by the authors, Licensee Intelligence Technology International Press. The article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA).