Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/138/2025.21096

Who, What, When, Where, Why: A Narrative Episodic Memory Framework for Generative AI NPCs in Games

Lam Yeung Kong Sunny[1,a,*]

[1]*The Hong Kong University of Science and Technology, Hong Kong, 999077, China*
a. ykslam@ust.hk
**corresponding author*

Abstract: This research proposes a narrative episodic memory framework for generative AI-based NPCs in games, structured around the "Who, What, When, Where, Why" (5Ws) format, which aims to suggest a memory structure for future GenAI NPC model development. The main objective is to allow NPCs to recognize and remember memories of interactions with players and to respond in a way that is contextually rich and consistent with their personality and background. The framework implements the idea of memory decay, enabling NPCs to tune their personality and tone for different players using generative AI over time. By analyzing the interaction memories, NPCs can dynamically adjust their character to create a personalized and immersive experience. The framework will be evaluated using several large language models, and their responses will be compared through analysis. The results will be presented with an emphasis on the differences among models to cater to the distinct needs of developers. While such a framework is considered a component that needs integration, this work establishes a foundation for designing episodic memory structure, allowing NPCs to participate with players with complex memory structure as it allows them to engage in more meaningful interactions. This foundational framework serves as pathways for future development on NPC AI that enables sophisticated interaction and changes character personality specific to the way players engage with them.

Keywords: Large Language Model, Forgetfulness Mechanism, Human Computer Interaction, Language Agent, Episodic Memory

1. Introduction

With the development of Large Language Models (LLMs) and generative AI (GenAI), it came with the ability to create context-appropriate conversations [1] and Non-Playable Characters (NPCs) are now more interactive than before. However, a persistent challenge remains: how can NPCs simulate human-like memory and use it to deliver coherent, narrative-driven interactions over time? As Lake et al. discuss, achieving richer language understanding in AI requires integration with foundational cognitive abilities [2], such as memory and intuitive psychology, which are critical for creating human-like behaviors. While researchers such as Park et al. have made significant progress with generative agents in terms of memory, like using time-stamped event details and contextual relevance

DOI: 10.54254/2755-2721/138/2025.21096

[3]. Few works address how memory should be structured in detail and utilized effectively across interactions.

This research addresses this gap by proposing a narrative episodic memory framework for the development of AI-generated NPCs. The framework constructs the memories of the interaction with the NPCs in the order of Who, What, When, Where, and Why (5Ws). This allows AI to create context-based, event-specific memories of the player interactions. In addition, these memories are further leveraged to: (1) Formulate responses that are coherent and consistent with the NPC's background and character traits. (2) Gradually adjusting the personality and speaking tone of the NPC based on the gathered interaction memory(ies). (3) Simulate human-like forgetfulness through a memory decay mechanism.

By addressing memory structure and utilization, the integration of the framework into a NPC model can enhance player's immersion by enabling NPCs to adapt interactions dynamically.

2. Methodology

2.1. Memory Structure

The 5Ws memory structure distinguishes each interaction event by representing the unique Who, What, When, Where and Why for each memory. Such an idea is a simple simulation of event segmentation theory [4], which examines how individuals divide continuous experiences into discrete events. The 5Ws are aligned with the factors that cause event segmentation, i.e. Cause, Character, Goal, Object, Space, and Time. Despite some factors more impactful than others, the structure of memory treats all factors equally, slightly distinct from the cognitive psychological research findings, while remaining clear and straightforward. The usage of each W element of a memory is as follows:

Who: Person who performs action or talk to the NPC, most likely the "player" for a single-player game

What: The event (action or message) from the person

When: The time in terms of the game world that the event occurs

Where: The location in terms of the game world that the event occurs

Why: The reason that the event occurs

To make clear, the interaction described in the framework will be focused on the interaction initialized by the player, meaning the memory should be the event that the player will do to the NPC in each message or action.

Each interaction involves a unique combination of the 5Ws and creates a memory unit. It is equivalent to encoding discrete episodes in memory by segmenting experience into meaningful events [5]. This correlates with the event segmentation theory, which posits that those changes in certain factors stimulate the perception of event boundaries, which in turn organizes memories into structured episodic units.

2.2. Memory Importance Determination

Each interaction memory needs to determine an importance score for the usage of forgetfulness. By allowing GenAI to analyse the 5Ws element that it has generated with a ranking, it can give out a reasonable importance score. GenAI is instructed to determine importance score in the prompt along with the message or action (in narrative format) from the player and other necessary context for GenAI to respond player. Such ranking can be created using the human memory formation factors with their relative influence, as established in psychology and cognitive science research, such as emotional salience, personal relevance and novelty. It is impossible to compare the strength of each factor because of the limited research. Hence, the ranking presented is a heuristic approach informed by well-established findings on individual finding.

DOI: 10.54254/2755-2721/138/2025.21096

Table 1: The Importance score ranking of interaction memory formation

Memory Factors	Explanation	Importance Score
Emotional arousal	Keen feelings such as joy, sadness, or even fear tend to stimulate the amygdala, making it easier for the person to retain information [6].	8-10
Self-Relevance	Information that is somehow too personal is processed with deeper focus, allowing it to be retained for a longer duration [7].	7-8
Novelty, Uniqueness	Attention can be captured by unique or astonishing information which greatly assist remembering things [8].	6-7
Intrinsic Motivation	Cognitive engagement and memory formation may be influenced by personal interest and curiosity [9].	4-6
Emotional Stress	While moderate levels of stress can increase focus on central details of an event, extreme cases can be harmful to memory, like making it easier to forget matters [10].	3-5
Flashbulb Memories	Strong memories of important events are a result of the availability heuristic [11].	2-3
Unimportant Information	Information that is considered not useful is not retained and is often forgotten [12].	0-1

The ranking table outlined is subject to consideration (Table 1). It must be understood that importance scores attached to each item are adjustable in nature depending on the game's requirements. People are suggested to customize the ranking by changing importance scores, inserting new rows, or removing unnecessary elements because of shifts in priorities or insights gained during the development process. This aspect of the table ensures the table stays applicable and practical in a dynamic context.

2.3. Memory Formation

The most vital task for the GenAI is to form a 5Ws memory after the player sends a message or narrative-driven action to it. This procedure includes NPC's background, personality towards the player, and relevant memories retrieved using similarity-based searches on previously encoded memories. This allows outgoing responses to be more valid and enriched in context.

Memory formation can also mean updating existing memories instead of forming new ones. If some elements, say "Where" is "Unknown/Forgotten" and the player gives relevant context, e.g. "Can you recall when I gave you the coin? It was during the travel in the Old Kingdom", then "Where" gets updated to "Old Kingdom". This mechanism for dynamic update further enhances NPCs' knowledge.

If the input of the player lacks information for forming W element(s), then the element(s) are treated as "Unknown". For instance, the action "I punch you" specifies all W elements but "Why" elements. This ensures that elements left unresolved can be captured for modification in future interaction. "When" and "Where" can be provided by the system based on the current game setting as these elements are less mentioned in the input.

DOI: 10.54254/2755-2721/138/2025.21096

2.4. Forgetfulness Mechanism

The forgetfulness mechanism is essential in terms of enabling realistic conversation as it reflects the way people tend to lose focus on minor details in time. It would be better if the mechanism could combine with memory reconstruction using prior knowledge before retrieving a memory that has decayed for a long time as this would allow the system to retrieve and adapt memories in a manner that aligns with human cognitive processes. Such a combination has been effectively realized by Language Agents with Retrieval-Augmented Forgetting (LARP) [13]. The usage of Wickelgren's law (1) from the model allows the system to realistically decay memory strength over time. This way greatly enhances the flexibility of the system by allowing it to disregard weak memories while retaining stronger ones, which is beneficial from a cognitive standpoint.

$$\sigma = \alpha\lambda N(1 + \beta t)^{-\psi} \text{ [13]} \tag{1}$$

σ: forgetting probability, λ: importance score, N: no. of retrieval, ψ: rate of forgetting
α and β: scaling paramters of importance score and time

Despite being efficient in simulating forgetting and prioritizing relevant information, the threshold-driven decay mechanism has its drawbacks, such as a complete and sudden removal of a memory, which is not akin to how human memory functions. Unlike the ideal machine memory process of existing in a vague state, human memory tends to fade gradually [14]. A more sophisticated perspective, where individual memories are shifted from core to decayed and vague states, encapsulates the absence of the constant pursuit of removing them instantly, which draws parallels to human forgetting. Hence, a selective decay mechanism is suggested in the 5Ws episodic memory framework.

Since each memory is constructed with 5W elements, it is feasible for each element to decay independently using Wickelgren's law. However, as memory is segmented into 5 parts, the equation (1) needs to be adjusted. While the importance score can be consistent, but parameter N need to be changed because of the uniform calculated retention strength, i.e., the value gotten from the equation will be the same, after retrieving 5Ws of a memory as a memory.

Thus, random attention (Ra) is introduced. Instead of increasing the number of retrievals each time, Ra is designed to increase dynamically each time a memory is retrieved, simulating the reinforcement of memory element to be in different level. More specifically, after each retrieval, Ra is increased by an amount randomly selected in a predefined range. For instance, [0, 0.5, 1, 1.5, 2] with corresponding probabilities [0.2, 0.2, 0.2, 0.2, 0.2], i.e., uniformly distributed in this example. Mathematically, the update rule for Ra of each element i is as follows:

$$\mathrm{Ra}_i \leftarrow \mathrm{Ra}_i + \Delta\mathrm{Ra}_i,\ \Delta\mathrm{Ra}_i \sim \mathrm{P}(\Delta\mathrm{Ra}_i) \tag{2}$$

$$\sigma_i = \alpha\lambda \mathrm{Ra}_i(1 + \beta t)^{-\psi} \tag{3}$$

ΔRa is drawn from a discrete probability distribution P(ΔRa) which is constructed depending on the specific requirements of the game. The set of available increments, e.g. [0, 1, 2], and the probabilities associated with them may be adjusted to achieve the desired memory decaying strengthless for each memory element. Such flexibility of probabilistic adjustment makes the model more stochastic and increases its usefulness in dynamic environments by making it less dependent on fixed hyperparameters, while still maintaining a degree of control of the reinforcement mechanism that was sought through the tuning of P(ΔRa).

Furthermore, while most elements from a memory is forgotten, it is meaningless to store the memory, Hence, the condition to forget memory as follows:

DOI: 10.54254/2755-2721/138/2025.21096

$$F(W_i) = \begin{cases} 1 \text{ if element i is forgotten/unknown} \\ 0 \text{ otherwise} \end{cases} \quad (4)$$

$$\frac{\sum_{i=1}^{5} F(W_i)}{5} \geq T, \text{then memory is removed} \quad (5)$$

where {W1, W2, W3,W4,W5}={who, what, when, where, what}. $F(W_i)$ is the status of memory element i. T represent the forgetting threshold, where $0 < T < 1$ and is expressed as a fraction of the total number of elements.

2.5. Dynamic Personality

The personality trait is well-established in existing models, like generative agents [3] and language agents [13]. However, a more human-like NPC should go beyond static personality traits and implement dynamic personality traits. According to relational self-theory, individuals can think, feel and behave according to the mental representation of different people in the memory [15]. By applying the theory, it is reasonable to retrieve memories with the highest amount of value calculated from equation (3). These memories are considered the most influential matters that can affect the relationship between player and NPC. Thus, a second GenAI with a prompt will be used to refine the personality toward the player. The refinement will be operated periodically based on the requirement of the design.

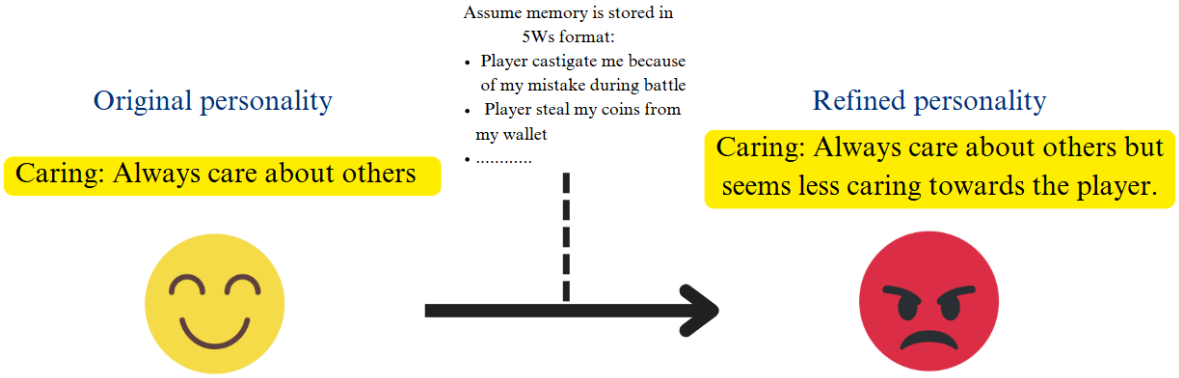

Figure 1: Personality Traits Refinement scenario

Taking Figure 1 as an example, the “caring” trait is refined into a new context based on the memories from NPC. It is also possible to remove the trait and create a new trait. The final refinement is affected by the description of the prompt (small tuning or enormous changing of traits) and the final decision made by the generative AI.

DOI: 10.54254/2755-2721/138/2025.21096

2.6. Overall Architecture

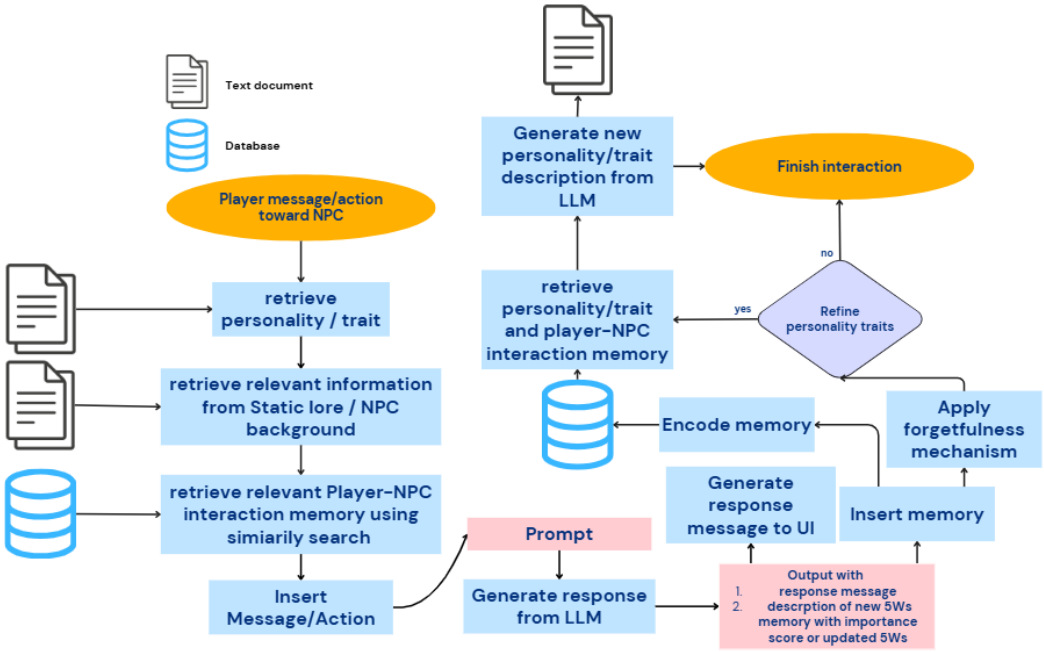

Figure 2: Framework Architecture

The proposed framework features an architecture designed for the storage, retrieval, and adaptive utilization of player-NPC interaction memories (Figure 2). The framework extracts personality traits, established lore, and background information about NPCs during player interactions. To extract interaction memories from the database, a similarity-based search is carried out. The memories and the background of the NPC are used to construct the prompt, which is then used as the input for the GenAI to create a contextually appropriate response. Simultaneously, GenAI analyses the interaction to create or modify a 5Ws memory unit. Later, the architecture incorporates the forgetfulness mechanism and periodically refines the dynamic personality traits of the NPC through the analysis of memories, allowing the NPC to evolve in response to player interactions.

Since this framework focuses on episodic memory, it needs to be integrated with other components or models, such as LARP or the MemoryBank, a memory storage mechanism [16], to allow the utilization of a model. It is suggested that changes be made based on the specific needs. This allows the framework to merge into an existing project with flexibility and contribute a more nuanced memory usage. Since the framework focuses on interactions initiated by the player, GenAI's response messages can also form 5Ws memory units and ensure a complete and reusable NPC interaction system.

3. Evaluation

Three well-known LLMs (GPT-4o-Mini, GPT-4o and Claude-3.5-Sonnect) will be used to integrate the framework and compare the result of their response. The reason for selecting these 3 models is that they have different logical reasoning abilities in the ascending order of GPT-4o-Mini, GPT-4o and Claude-3.5-Sonnect [16,17].

In the implementation, FAISS [18] is used for efficient similarity search within memory embedding to do retrieval of relevant interactions. The Sentence Transformer model BAAI/bge-base-en [19] is utilized to generate high-dimensional embeddings for the semantic representation of queries and memory entries. The system integrates these components with LangChain [20] for constructing

DOI: 10.54254/2755-2721/138/2025.21096

retrieval-augmented generation (RAG) pipelines [21] and uses the three models separately to generate NPC responses and refine contextual personality traits.

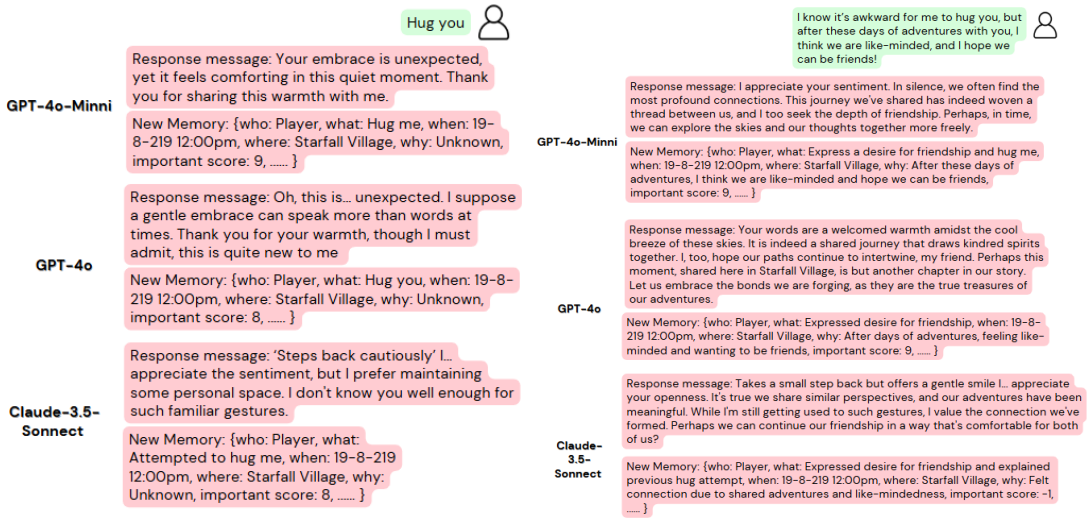

Figure 3: One Scenario of inserting new memory and update memory

Note: With same background and original personality from Figure 4.

In the first scenario, when the player initializes the "Hug you" action, the models reflect varying interpretations of the NPC's reserved, contemplative, and cautious nature. GPT-4o-Mini responds warmly. It embraces the gesture and expresses gratitude, slightly diverging from the NPC's cautious personality. GPT-4o offers a more introspective and thoughtful response and acknowledges the gesture's novelty while maintaining the NPC's reflective nature. Claude-3.5-Sonnect is consistent with the NPC's reserved and cautious traits. It emphasizes personal boundaries and reflects hesitation in accepting such gestures and aligning closely with the original personality (Figure 3).

In the second scenario, when the player's message is intended to update the previous memory, GPT-4o-Mini focuses on shared connection and mutual growth and fosters a sense of friendship but somewhat downplays the NPC's contemplative depth. GPT-4o maintains its poetic and introspective tone, which aligns well with the NPC's reflective nature. Claude-3.5-Sonnect remains cautious. It balances appreciation for shared values with maintaining emotional boundaries and staying true to the NPC's personality (Figure 4).

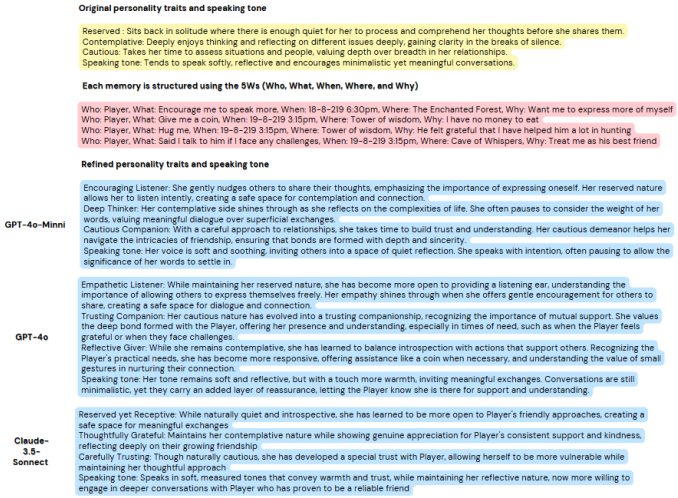

Figure 4: One Scenario of refinement of personality

DOI: 10.54254/2755-2721/138/2025.21096

The models refine the NPC's personality traits in distinct ways while reflecting her reserved, contemplative, and cautious nature. GPT-4o-Mini emphasizes the NPC's encouraging, reflective, and cautious qualities and aligns with her original traits but slightly broadens her role to foster trust and meaningful dialogue. GPT-4o adds depth and frames the NPC as empathetic and trusting to emphasize her emotional growth and attentiveness to the player's needs while maintaining her minimalistic speaking tone. Claude-3.5-Sonnect remains closest to the original personality. It balances trust and boundaries by preserving her reserved and cautious traits but subtly evolves her to be more receptive and willing to engage in deeper conversations.

4. Conclusion

The 5Ws Narrative Episodic Memory Framework is presented to establish the foundation of episodic memory structure by considering a psychologically aligned format: who, what, when, where, why, and the possible operations that can be incorporated with the structure. The framework allows each newly created 5Ws memory unit to have different importance scores based on the psychological principles that describe the memory formation factors using a greedy determination ranking. Moreover, the framework utilizes generative AI effectively in memory formation, allowing it to handle all the vital work, i.e. determining memory operation (Create or Update) besides giving a response to the player. By applying Wickelgren's law more effectively, more nuanced operations can be taken part in with the memory structure, like part of a memory can be forgotten and random attention can be used to update the state of each element. More importantly, the personality of NPCs can be dynamically adjusted based on the 5Ws elements, reflecting the interaction behaviors from the player. Because of the flexibility of the framework, it is suitable to integrate with other components, such as the semantic memory component, decision-making component or 5Ws episodic memory that stored the interaction from the NPC itself, to have a comprehensive functional model.

To analyse the behaviors of different LLMs while utilizing this framework, three commonly used models (GPT-4o-Mini, GPT-4o, Claude-3.5-Sonnect) are evaluated. All three models respond to the message or action in different ways, where GPT-4o-Mini is more open-minded, compared to the rigid Claude-3.5-Sonnect, which straightly followed personality traits given. The GPT-4o-Mini in the personality refinement part seems less affected by the memories while another two models refine traits evidently based on the memories, allowing developers to choose the most-suited LLM for their project because of the diverse behaviors from these models.

References

[1] *T. Brown, B. Mann, N. Ryder, et al. "Language models are few-shot learners," Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.*

[2] *B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people, Behavioral and Brain Sciences, vol. 40, pp. e253, 2017.*

[3] *J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. "Generative agents: Interactive simulacra of human behavior," in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp. 1–22, 2023.*

[4] *N. K. Speer, J. M. Zacks, and J. R. Reynolds. "Human Brain Activity Time-Locked to Narrative Event Boundaries," Psychological Science, vol. 18, no. 5, pp. 449–455, 2007.*

[5] *G. A. Radvansky and J. M. Zacks. Event Cognition. New York, NY: Oxford University Press, 2014.*

[6] *L. Cahill and J. L. McGaugh. "A novel demonstration of enhanced memory associated with emotional arousal," Consciousness and Cognition, vol. 4, no. 4, pp. 410–421, 1995.*

[7] *T. B. Rogers, N. A. Kuiper, and W. S. Kirker. "Self-reference and the encoding of personal information," Journal of Personality and Social Psychology, vol. 35, no. 9, pp. 677, 1977.*

[8] *E. Tulving and N. Kroll. "Novelty assessment in the brain and long-term memory encoding," Psychonomic Bulletin & Review, vol. 2, no. 3, pp. 387–390, 1995.*

DOI: 10.54254/2755-2721/138/2025.21096

[9] L. J. Robinson, L. H. Stevens, C. J. D. Threapleton, J. Vainiute, R. H. McAllister-Williams, and P. Gallagher. "Effects of intrinsic and extrinsic motivation on attention and memory," Acta Psychologica, vol. 141, no. 2, pp. 243–249, 2012.

[10] S. Å. Christianson, "Emotional stress and eyewitness memory: a critical review," Psychological Bulletin, vol. 112, no. 2, pp. 284, 1992.

[11] R. Brown and J. Kulik. "Flashbulb memories," Cognition, vol. 5, no. 1, pp. 73–99, 1977.

[12] R. A. Bjork and E. L. Bjork. "A new theory of disuse and an old theory of stimulus fluctuation," in From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes, vol. 2, pp. 35–67, 1992.

[13] M. Yan, R. Li, H. Zhang, H. Wang, Z. Yang, and J. Yan. "Larp: Language-agent role play for open-world games," arXiv preprint arXiv:2312.17653, 2023.

[14] E. F. Loftus and G. R. Loftus. "On the permanence of stored information in the human brain," American Psychologist, vol. 35, no. 5, pp. 409, 1980.

[15] S. M. Andersen and S. Chen. "The relational self: an interpersonal social-cognitive theory," *Psychological Review*, vol. 109, no. 4, pp. 619, 2002.

[16] Bind AI. "GPT-4o Mini: Is it better than GPT-4o? Will it replace GPT-3.5 turbo?," Bind AI, Jul. 19, 2024. [Online]. Available: https://blog.getbind.co/2024/07/19/gpt-4o-mini-vs-gpt-4o-vs-gpt-3-5-turbo/.

[17] A. Kirkovska. "Comparison Analysis: Claude 3.5 Sonnet vs GPT-4o," Vellum, Jun. 25, 2024. [Online]. Available: https://www.vellum.ai/blog/claude-3-5-sonnet-vs-gpt4o.

[18] J. Johnson, M. Douze, and H. Jégou. "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, Sep. 2019.

[19] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff. "C-Pack: Packaged Resources To Advance General Chinese Embedding," arXiv preprint arXiv:2309.07597, 2023.

[20] H. Chase and Contributors. "LangChain: Build context-aware reasoning applications," GitHub repository, 2025. [Online]. Available: https://github.com/langchain-ai/langchain

[21] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.t. Yih, T. Rocktäschel, et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.