Human–Computer Interaction Optimization through LLM-Based Cognitive Modeling and Personalized Dialogue Memory

Blake A. Ortiz
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
blake1975@uc.edu

Olivier Robles
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
robles599@binghamton.edu

Abstract

The rapid evolution of large language models has transformed human–computer interaction, yet persistent challenges remain in aligning system behavior with user intent, context, and long-term preferences. This paper presents a system-level examination of how cognitive modeling and personalized dialogue memory can be integrated into LLM-based interactive systems to optimize human–computer interaction. We argue that treating interaction as a continuous cognitive process, rather than a series of stateless exchanges, requires architectures that embed working memory, episodic recall, and user-specific semantic representations directly into the inference loop. By analyzing structural trade-offs across centralized and federated memory designs, latency-sensitive retrieval mechanisms, and adaptive personalization engines, the paper maps the design space for scalable, robust, and fair cognitive dialogue systems. We explore the interplay between memory decay, forgetting policies, and the right to be forgotten under governance frameworks such as the GDPR, revealing tensions between personalization depth and data minimization. Through case illustrations in healthcare and enterprise deployment, we show how architectural decisions around memory persistence, knowledge conflict resolution, and model retraining frequency directly impact fairness, robustness, and sustainability. The discussion extends to infrastructure implications, including on-device versus cloud-based processing, the carbon footprint of lifelong memory stores, and the long-term socio-technical risks of feedback loops that amplify cognitive biases. The paper concludes by identifying actionable policy pathways and design principles that guide the responsible deployment of LLM-based cognitive interaction systems, highlighting future research directions at the intersection of human-centered AI and systems engineering.

Keywords

Human-computer interaction, large language models, cognitive modeling, personalized memory, dialogue systems, system architecture, fairness, governance.

1. Introduction

Contemporary human–computer interaction (HCI) is undergoing a paradigmatic shift driven by the integration of large language models (LLMs) into conversational agents, productivity tools, and decision-support systems. The foundational transformer architecture [1] has enabled language generation of unprecedented fluency and coherence, while the broader

category of foundation models [2] has extended these capabilities across multimodal and task-agnostic settings. When augmented with retrieval mechanisms [3] and fine-tuned through human feedback [4], LLMs can produce contextually relevant and instructionally aligned outputs that approach human-level conversational quality. Yet, despite these advances, a critical limitation persists: most deployed systems treat each interaction as a largely isolated event, lacking the depth of personalized, temporally extended cognitive engagement that characterizes effective human communication. Recent proposals to leverage LLM-based user memory systems for query matching and intent modeling [5] mark the beginning of a shift toward more persistent and adaptive interaction frameworks, but the full spectrum of system-level implications remains underexplored.

The optimization of HCI through cognitive modeling and personalized dialogue memory requires a holistic understanding of how memory architectures, personalization strategies, and infrastructure choices intersect with human values, governance norms, and sustainability goals. Unlike traditional dialogue systems that rely on slot-filling or finite-state grammars, LLM-based cognitive systems must reconcile multiple timescales of memory, ranging from immediate working memory for turn-by-turn coherence to long-term episodic and semantic stores that shape user identity and evolving preferences. This paper frames the challenge as a systems design problem in which structural trade-offs—between privacy and personalization, latency and richness, fairness and adaptation, cost and recall fidelity—define the feasible operational envelope. The subsequent sections examine these trade-offs from a cross-disciplinary perspective, integrating insights from cognitive architecture research, memory-augmented neural networks, responsible AI governance, and large-scale system deployment.

The paper's contributions center on a comprehensive system-level analysis rather than a novel algorithmic contribution. We synthesize design patterns for cognitive dialogue memory, evaluate personalization paradigms under fairness constraints, and articulate governance and infrastructure roadmaps. Case illustrations grounded in healthcare conversational agents and enterprise knowledge assistants are employed to concretize abstract trade-offs, demonstrating, for example, how memory retention policies can mitigate bias amplification and how edge-cloud partitioning can support both low-latency interactions and data sovereignty. By constructing a narrative that bridges engineering architecture and socio-technical policy, the paper provides researchers and practitioners with an integrative view of what it means to optimize human–computer interaction through cognitive LLM systems.

2. Background and Foundational Concepts

The design of LLM-based systems that emulate cognitive processes and maintain personalized dialogue memory rests on three intersecting research streams: cognitive architectures and their modern reinterpretations through neural models, memory-augmented dialogue systems with personalization capabilities, and established principles for human-AI interaction optimization. This section synthesizes these streams to establish the conceptual grounding for the architectural analysis that follows.

2.1 Cognitive Architectures and Large Language Model Integration

Classical cognitive architectures, such as ACT-R and Soar, were developed to model human cognition through modular subsystems for declarative and procedural memory, goal management, and perceptual-motor interfaces. An extensive survey of forty years of cognitive architectures [16] highlights the enduring relevance of these systems in accounting for learning, reasoning, and adaptive behavior. While earlier symbolic architectures struggled

with the scalability required for open-domain language interaction, the training dynamics of LLMs exhibit emergent properties that partially recapitulate human-like cognitive operations, including analogical reasoning and context-sensitive inference [17]. The core challenge addressed in this paper resides in bridging the gap between the imprecise memory and reasoning capabilities of LLMs and the structured, persistent, and verifiable memory required for sustained personalized interaction. Cognitive science research on human learning suggests that machines that learn and think like people must incorporate causal models, intuitive physics, and compositional generalization [8], none of which emerge automatically from next-token prediction objectives. Integrating LLMs within a cognitive modeling framework therefore demands architectural scaffolding that explicitly manages working memory buffers, long-term storage, and retrieval policy logic, all while remaining computationally tractable at scale.

2.2 Personalized Dialogue and Memory Systems

Memory networks were among the first neural architectures to demonstrate how explicit storage and iterative read operations could support question answering and reasoning over facts [6]. The subsequent end-to-end memory network formulation [7] further showed that such mechanisms could be trained jointly with the rest of the model, removing the need for hard-coded retrieval logic. In dialogue systems, personalized response generation has been advanced through persona-aware models that condition on user profile embeddings [10] and through diversified memory structures that store persona-related facts to maintain consistency across multiple conversational turns [9]. These approaches, however, often treat personalization as a static set of attributes, neglecting the continuous evolution of user preferences, emotional state, and knowledge. True cognitive personalization requires a dynamic memory that encodes not only what the user has explicitly stated but also inferred intents, interaction style, and gradually shifting belief states. Moreover, the memory must interact bidirectionally with the generative process, allowing the system to selectively attend to relevant past episodes while also updating its internal representations based on ongoing interaction. This paper extends the notion of personalized dialogue memory by situating it within a larger cognitive architecture that governs when and how memories are consolidated, retrieved, or forgotten.

2.3 Human–Computer Interaction Optimization Principles

The optimization of HCI through AI has long been guided by principles that emphasize transparency, controllability, and alignment with user mental models. The guidelines for human-AI interaction put forth by Amershi et al. [11] provide a set of design recommendations spanning initial user onboarding, error recovery, and adaptive personalization. In the context of LLM-based cognitive systems, these principles must be reinterpreted when the interaction model shifts from deterministic command-response patterns to fluid, open-ended dialogues that learn and adapt over months or years. Users may no longer be able to fully introspect the system's memory state or understand why a certain personalized response diverges from their past experience, which raises novel challenges for explainability and trust. Furthermore, the heavy reliance on user data for personalization invokes fairness concerns: if personalization disproportionately benefits certain demographic groups due to uneven data quantity or quality, the interaction experience can become inequitable. These issues foreground the need for governance mechanisms and fairness-aware design, which we analyze in depth in later sections.

3. System Architecture for LLM-Based Cognitive Modeling and Dialogue Memory

A robust system architecture for cognitive dialogue optimization must orchestrate the interplay between language generation, memory operations, and personalization logic across multiple timescales. We decompose the architecture into four interconnected layers: the base LLM inference core, the cognitive memory manager, the personalization engine, and the governance-and-monitoring overlay. Each layer presents distinct structural trade-offs that influence scalability, latency, privacy, and fairness.

The base LLM inference core serves as the primary language generator and zero-shot reasoner. While one architectural extreme places all cognitive and memory functions inside a monolithic model fine-tuned to internalize personalized context through prompt engineering or parameter-efficient adaptation [18], this approach collapses multiple concerns into a single black box, complicating auditability and modular updates. A more transparent alternative separates memory and personalization from the core model, routing the LLM's attention over a managed memory store. This separation permits independent scaling of the memory subsystem, the injection of governance hooks for data access logging and consent enforcement, and the ability to swap or retrofit the LLM without disturbing accumulated user memories. However, the indirection adds latency for each memory read, which becomes particularly acute in real-time voice interfaces that tolerate end-to-end latencies below 300 milliseconds. Architectures can mitigate this by using a two-tier memory: a fast, cache-resident working memory that holds recent turns and active user intents, and a slower, persistent long-term store indexed by a vector database that supports semantic retrieval of relevant episodes. The trade-off involves freshness of recalled memories: frequent synchronization with the long-term store ensures high fidelity but consumes compute and I/O, while aggressive caching risks staleness that degrades personalization quality, notably in rapidly evolving task contexts like travel planning or crisis counseling.

The cognitive memory manager implements the logic for encoding, retrieval, consolidation, and forgetting. It draws inspiration from the multi-component working memory model and the complementary learning systems theory of hippocampus-neocortex interaction. A structural tension emerges between storing raw dialogue logs and maintaining a structured, symbolic summary: raw logs preserve maximum information for future retrieval but dramatically inflate storage costs and retrieval latency over months of interaction; compressed summaries protect user privacy and reduce footprint but may irreversibly discard nuances that later become crucial for accurate intent modeling. Systems designers must therefore devise adaptive summarization policies that balance compression ratio with retention of affect, uncertainty, and subtle preferences. For instance, a healthcare coaching agent may need to preserve detailed records of symptom descriptions with temporal markers while abstracting away irrelevant small talk, requiring a context-sensitive filtering mechanism that itself may be realized as a smaller language model trained on privacy-preserving summarization directives.

The personalization engine translates memory representations into action: it adjusts response style, recommends content, and proactively surfaces information. This layer often implements a multi-armed bandit or reinforcement learning policy that continuously optimizes interaction outcomes such as task completion rate or user satisfaction, drawing on memory as a state representation. An important structural decision is whether the personalization model is trained globally from aggregated user data and then locally fine-tuned, or whether personalization is entirely on-device with federated learning loops. Centralized personalization benefits from massive data and can discover latent preference clusters, but it introduces systemic risk if a few dominant user profiles bias the global model, leading to

homogenization that suppresses minority interaction styles. Federated architectures preserve user sovereignty and align with regulations such as GDPR that emphasize data minimization, yet they introduce new complexities in coordinating model updates when memory schemas diverge across devices. Furthermore, the choice of personalization granularity—segmenting user groups versus per-user modeling—has direct implications for fairness: overly granular personalization can amplify historic biases embedded in an individual's interaction history, while coarse segmentation may fail to capture legitimate individual needs.

The governance-and-monitoring overlay cuts across all layers to enforce fairness, safety, and policy compliance. It intercepts memory read and write operations to apply differential privacy guarantees, audit trails, and contentious content filters. For example, right-to-be-forgotten requests mandate the targeted deletion of specific memories without degrading global model parameters—a challenge that resembles machine unlearning. Architectural support for selective forgetting often requires a data structure that associates each memory fragment with a user-controlled metadata tag and enables efficient traceability through the training data lineage. This overlay must also measure and report fairness metrics, such as intent recognition parity across demographic subgroups, triggering alerts when personalization diverges beyond acceptable thresholds. While the overhead of continuous monitoring may appear to be a burden, it is more appropriately regarded as an essential architectural budget that safeguards long-term system trust and regulatory viability.

4. Personalization and Memory Management Paradigms

The core promise of personalized dialogue memory is that the system will evolve into a collaborative cognitive partner rather than a stateless tool. Achieving this vision necessitates deep consideration of how personalization is modeled, how conflicting information from multiple interactions is reconciled, and how the memory life cycle is governed to avoid cognitive overload and bias entrenchment.

Personalization paradigms range from explicit user modeling, where the system maintains a declarative profile that users can inspect and edit, to implicit, emergent personalization driven purely by interaction patterns. Explicit profiles offer transparency and user control, aligning closely with HCI guidelines on user agency [11], but they impose a setup burden and may fossilize if users do not actively update them. Implicit personalization, on the other hand, adapts seamlessly as the user's language, topical interests, and emotional tones shift over time, yet it risks creating an unknowable "shadow model" of the user that can persist outdated or inaccurate assumptions. A promising hybrid approach maintains a lightweight explicit skeleton—perhaps a handful of high-level user attributes or goals—while fleshing out the model with dense implicit representations learned from dialogue memory. The structural challenge is to design synchronization protocols so that changes in explicit settings propagate to implicit layers and prompt the re-evaluation of past inferences without retraining the entire personalization model from scratch.

Memory reconciliation becomes critical when the system must integrate contradictory signals from a user over time. A user may state a strong preference for vegetarian meals one week and later order a dish containing meat in a food assistant scenario. Naively overwriting the old memory with the new fact can cause erratic behavior, while retaining all contradictory facts without conflict resolution can lead to incoherent personalization that oscillates between preferences. Cognitive architectures inspired by human belief revision offer strategies such as confidence-weighted memory that decays older contradicting facts only gradually, or source-attributed memory that tags each fact with the context and certainty of its acquisition. This

tagging allows the personalization engine to reason about the relative reliability of different user utterances—for instance, giving higher weight to facts expressed as explicit preferences in a settings interaction than to those implicitly inferred from a single query. Moreover, memory management must support user-initiated correction mechanisms: when a user says "Forget that I ever liked that restaurant," the system should execute not only a surface-level deletion but also a re-examination of associated inferential chains that might have been built upon the retracted fact.

In long-running deployments, the unbounded growth of dialogue memory threatens both computational efficiency and cognitive relevance. Forgetting policies thus become a first-class design concern, not a failure mode. A common approach is to implement a time-based decay that gradually reduces the retrieval probability of older memories unless they are periodically reinforced, analogous to the spacing effect in human memory. Other systems employ importance-based pruning driven by a salience model that predicts the future utility of each memory, enabling the system to retain high-value personal knowledge while discarding ephemeral or redundant interactions. Both approaches must be carefully governed to avoid discriminatory forgetting that disproportionately removes content associated with underrepresented groups or non-standard interaction patterns. For instance, a salience model trained predominantly on data from native English speakers may underestimate the long-term relevance of code-switched or dialectal content, causing the memory system to systematically forget these linguistic expressions and degrade the interaction quality for such users. Fairness auditing of forgetting mechanisms thus becomes an integral element of personalization governance.

A healthcare setting provides a compelling illustration of these dynamics. Consider a cognitive dialogue companion designed to support patients managing chronic conditions such as diabetes. The system must maintain a detailed memory of medication schedules, dietary logs, and symptom timelines, while adapting to the patient's changing motivational state and health literacy level. The personalization engine may learn that a particular patient responds better to gentle encouragement framed as small daily goals, whereas another requires direct, data-centric feedback. If the memory system permanently records every dietary slip, however, it could inadvertently reinforce a negative self-image and discourage the user, raising ethical concerns. Here, forgetting policies must be tuned to fade minor deviations while preserving medically relevant trend lines, all within a regulatory framework that governs health data retention. The architectural separation of memory management from core generation enables domain-specific governance modules that encode clinical guidelines, audit data handling, and enforce compliance with standards such as HIPAA. This case highlights that the choice of forgetting granularity is not a purely technical optimization but a profoundly human-centered decision with clinical consequences.

5. Governance, Fairness, and Robustness

Bringing persistent cognitive memory into LLM-based interaction systems magnifies the importance of governance frameworks that encompass data stewardship, non-discrimination, and robustness to adversarial environments. The encoding of user identity and history within a mutable memory construct blurs the line between a tool and an intimate agent, making it necessary to define accountability structures that address memory misuse, bias accumulation, and failure modes.

Fairness in personalized systems demands that the benefits of adaptation be distributed equitably across diverse user populations. Empirical studies of language models have

demonstrated that training data biases can lead to performance disparities along axes of race, gender, and dialect [12], and memory systems that selectively recall or forget user history risk compounding these disparities. A personalization engine that primarily trains on interaction logs from high-engagement users may systematically underperform for users who interact less frequently or whose linguistic style diverges from the training majority. This is a structural fairness concern that cannot be fully remedied by simple post-hoc debiasing of the underlying LLM, because the feedback loop between memory retrieval and generation can amplify small initial differences over time. Robust governance requires the continuous monitoring of interaction quality metrics disaggregated by relevant demographic and behavioral segments, combined with a mechanism to rebalance memory retrieval weights when disparate impact is detected. Such mechanisms might temporarily boost the salience of stored memories for underserved groups until performance parity is restored, but they also introduce the ethical question of whether unequal treatment of memory retrieval is acceptable in pursuit of equality of outcome.

The right to be forgotten, codified in regulations such as GDPR, acquires new complexity within cognitive dialogue systems trained on evolving memory stores. Unlike static databases where deletion is a well-defined operation, a user's personal data in an LLM-based memory system may be encoded not only in explicit log entries but also in the latent weights of a fine-tuned personalization adapter. True compliance therefore demands machine unlearning techniques capable of erasing the influence of specific data points without a full and costly retraining cycle. Architectural choices such as storing personal embeddings in a separate, mutable vector store rather than baking them into the generative model parameters greatly simplify deletion, as the model itself remains impartial and only the retrieval index must be updated. However, this separation reintroduces the latency trade-off discussed earlier and may reduce the depth of personalization that can be achieved without co-adapting model weights. Governance frameworks will need to provide guidance on the acceptable granularity of forgetting, distinguishing between the removal of raw conversational data and the right to have one's inferred preferences and behavioral patterns expunged from the cognitive model.

Robustness challenges arise from both accidental noise and deliberate adversarial manipulation of the dialogue memory. A user might inadvertently provide contradictory or misleading information in the course of natural conversation, or a malicious third party might inject crafted prompts into a shared device session to poison the memory with false personal facts that steer future behavior. Memory systems must therefore incorporate integrity checks such as source provenance tracking, redundancy-based fact verification against trusted knowledge bases, and confidence thresholds that flag suspiciously rapid shifts in user profiles. The architectural insertion of a reasoning step that cross-references retrieved memories with a set of invariant user-safe constraints can prevent catastrophic personalization failures. In high-stakes domains like finance or legal advice, a memory recall should always be mediated by a policy layer that ensures compliance with regulated boundaries, overriding personalized inclinations when necessary. Designing such guardrails without undermining the fluidity and naturalness that define good HCI remains an open challenge, and one that benefits from a cognitive modeling perspective that treats the guardrail as analogous to the human executive control processes that inhibit impulsive responses.

Policy implications extend to the governance of cross-session and cross-platform memory sharing. As users expect their cognitive companion to seamlessly transition from a smart speaker to a vehicle interface to a desktop application, the memory system becomes a portable

user model that must be synchronized across devices and administrative domains. This portability creates tension between convenience and security: a unified memory cloud enables consistency but becomes a high-value target for data breaches, while fully local, device-scoped memory enhances privacy but leads to fragmented and inconsistent interaction experiences. Federated identity protocols and zero-knowledge proofs for memory access may offer a compromise, allowing the system to compute personalization signals without exposing raw memory contents. The development of industry-wide standards for memory portability, akin to the Data Transfer Project initiated by major technology firms, will be instrumental in enabling user-centric cognitive ecosystems without vendor lock-in.

6. Deployment and Infrastructure Considerations

Translating the architectural vision of a cognitive dialogue system into a production-grade deployment reveals additional layers of trade-offs that span compute resource allocation, networking, lifecycle management, and sustainability. The scale of modern LLMs, which can comprise hundreds of billions of parameters, imposes significant inference costs, and adding a rich memory retrieval pipeline further strains latency and throughput budgets. A typical architecture deployed in enterprise customer support must simultaneously serve thousands of concurrent users, each with a multi-year interaction history stored as high-dimensional vectors. Without careful engineering, the tail latency arising from memory lookups can negate the responsiveness that users expect from a conversational interface.

A principal structural decision concerns the partitioning of workloads between cloud infrastructure and edge devices. Edge-dominant architectures run a compressed, quantized model on the user's device and maintain a local memory store that is fast, private, and resilient to network interruptions. This model naturally aligns with personal data sovereignty but limits the depth of retrieval and the richness of the base LLM due to hardware constraints. Cloud-dominant deployments centralize heavy inference and large-scale memory indices, offering unmatched recall and personalization breadth at the cost of network dependency and potential exposure of sensitive data in transit or at rest. A pragmatic design blends both: an on-device memory module that handles short-term working memory and generates an anonymized intent embedding, which is then securely transmitted to a cloud-based long-term memory system for episodic retrieval. The segmentation of responsibilities demands robust fallback strategies so that the conversation degrades gracefully—switching to locally cached responses or generic templates—when connectivity is lost. This hybrid deployment model also allows the cloud component to aggregate anonymized memory access patterns for global model improvement without exposing raw personal content.

Lifecycle management encompasses the continuous training, evaluation, and updating of models and memory retrieval indices. Unlike traditional software, a cognitive dialogue system is never "finished"; it must adapt to shifting linguistic patterns, new domain knowledge, and evolving user expectations. MLOps pipelines must support incremental learning cycles in which the personalization engine updates its policy and the memory indexing adapts to new embedding spaces when the base LLM is versioned. A particularly acute challenge is the mismatch between the slow cadence of base model retraining and the rapid pace of personal memory change. When a new LLM release introduces a different latent representation space, all previously stored memory embeddings may become semantically misaligned, requiring a costly re-indexing operation. System architects can partially mitigate this through training regimes that regularize embedding spaces across model versions or by adopting an adapter-based architecture that decouples the personalization head from the evolving base model. The

choice of retraining frequency also has significant carbon footprint implications, as training massive models is energy-intensive. Organizations committed to environmental sustainability must weigh the performance gains of frequent retraining against the associated emissions, exploring techniques such as sparse update, efficient fine-tuning, and on-demand retraining triggered only by significant distributional drift.

The infrastructure must also support the logging and auditing mechanisms essential for governance. Every memory access, update, and deletion should generate tamper-proof log entries that can be used to answer user queries about what the system "remembers" and to comply with regulatory audits. These logging requirements impose storage and throughput overhead that grows with user base and interaction frequency. A tiered logging architecture, where high-resolution logs are retained only for a short window for debugging and a summary of semantically meaningful memory operations is archived for the long term, can balance accountability with cost. The deployment of a cognitive memory system thus becomes an exercise in managing competing operational requirements: delivering low-latency, highly personalized interactions while maintaining a transparent, auditable, and sustainable infrastructure footprint.

7. Sustainability and Long-Term Adaptation

Sustainability in cognitive dialogue systems extends beyond environmental concerns to encompass the socio-technical viability of systems that learn and change over years of operation. A system that accumulates user memory indefinitely may not only become computationally unwieldy but also impose a psychological burden on users who interact with an agent that never forgets—an effect that can erode the spontaneity and forgiveness inherent in natural human conversation. Designing for sustainability therefore requires a careful choreography of memory consolidation, graceful forgetting, and limits on personalization depth that respect human cognitive comfort zones.

From an environmental perspective, the ongoing storage and processing of dialogue memory at global scale has a non-trivial energy footprint. Large-scale vector databases that index billions of personal memories demand substantial memory and compute resources, and each retrieval across that index consumes electricity. Techniques such as quantized vector storage, approximate nearest neighbor search with bounded accuracy loss, and hierarchical caching can reduce this footprint, but they must be evaluated against the personalization cost. A policy-oriented view suggests that sustainability metrics—measured as carbon emissions per thousand meaningful interactions—should be standardized and reported alongside accuracy and latency figures. Such transparency would enable platform developers and regulators to set benchmarks for green cognitive interaction services and incentivize the adoption of energy-proportional architectures that scale down during idle periods.

Long-term adaptation is not only a technical challenge but a question of user trust and social embedding. A cognitive dialogue system that adapts too eagerly may appear fickle or intrusive, while one that adapts too slowly frustrates users who expect it to learn. The design of adaptation rates must therefore be informed by empirical HCI research rather than driven solely by optimization metrics. For example, a conversational agent deployed in educational settings for children might benefit from rapid adaptation to keep pace with developing language skills, yet it must simultaneously retain stable pedagogical scaffolding. This suggests a multi-timescale personalization model where surface-level interaction style can adapt quickly, while deeper user understanding—such as a learner's fundamental misconceptions—evolves more slowly and only after multiple reinforcing observations. Such

a model mirrors human social cognition, where first impressions form rapidly but core relationship models change only through sustained experience. Embedding these human-inspired rhythms into memory policies can yield interactions that feel more natural and reduce the uncanny valley effect that sometimes characterizes highly optimized but socially unnatural AI.

Policy incentives will play a central role in shaping sustainable deployment. Public funding for research on green AI and regulatory carrots for companies that adopt open standards for memory interoperability and deletion can accelerate the transition toward cognitive systems that are both powerful and environmentally responsible. Furthermore, as cognitive dialogue agents become more deeply woven into healthcare, legal, and civic services, governments may mandate that these systems be designed with a "sustainability by design" principle that includes the right to digital forgetting, algorithmic impact assessments that cover long-term societal effects, and caps on per-capita resource consumption for personalized AI services. These measures can prevent a tragedy of the commons in which competitive pressure drives ever-increasing personalization granularity and memory depth at the expense of collective sustainability and fairness.

8. Conclusion

This paper has provided a system-level analysis of how LLM-based cognitive modeling and personalized dialogue memory can be harnessed to optimize human–computer interaction. By investigating the architectural decomposition into inference, memory management, personalization, and governance layers, we have illuminated a rich design space where structural decisions about centralization, forgetting, retrieval, and monitoring reverberate across performance, privacy, fairness, and sustainability dimensions. We have argued that treating dialogue memory as a continuous cognitive construct—rather than a simple log—enables more natural, adaptive, and empowering interactions, yet it simultaneously surfaces profound challenges that cannot be resolved through engineering optimization alone.

The case studies of healthcare companions and enterprise assistants demonstrated that forgetting policies, memory reconciliation strategies, and deployment architectures are deeply entangled with ethical and regulatory imperatives. Fairness-aware memory design emerges as a critical requirement to prevent personalization from amplifying social inequities, while governance mechanisms ensuring the right to be forgotten and auditable memory trails provide the necessary trust infrastructure for long-term human–AI partnerships. On the operational front, edge-cloud hybrid architectures offer a pragmatic path to balancing low-latency interaction and data sovereignty, but they demand continued innovation in model compression, federated learning, and energy-efficient retrieval.

Looking ahead, research must move beyond isolated component optimization to embrace whole-system co-design, where cognitive modeling principles, memory semantics, interaction guidelines, and sustainability metrics are jointly engineered from the outset. The development of standardized benchmarks for cognitive dialogue performance—incorporating not only accuracy but also fairness, forgetting fidelity, and carbon intensity—would enable comparative evaluation and drive the field toward more responsible systems. Ultimately, the optimization of human–computer interaction through personalized cognitive memory is not merely a technical frontier; it is a socio-technical project that requires thoughtful collaboration across disciplines to shape a future in which AI systems genuinely augment human intellect and agency without compromising the values we hold dear.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

2. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

4. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

5. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.

6. Weston, J., Chopra, S., & Bordes, A. (2015). Memory networks. In Proceedings of the 3rd International Conference on Learning Representations (ICLR).

7. Sukhbaatar, S., Weston, J., & Fergus, R. (2015). End-to-end memory networks. Advances in Neural Information Processing Systems, 28.

8. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253.

9. Song, H., Wang, Y., Zhang, W., Liu, X., & Chua, T.-S. (2019). Exploiting persona information for diverse and personalized dialogue generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2490–2501).

10. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 2204–2213).

11. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13).

12. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623).

13. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59–68).

14. Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint arXiv:2102.02503.

15. Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), 1–29.

16. Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. Artificial Intelligence Review, 53(1), 17–94.

17. Clark, P., Tafjord, O., & Richardson, K. (2020). Transformers as soft reasoners over language. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (pp. 3882–3888).

18. Zaken, E. B., Ravfogel, S., & Goldberg, Y. (2022). BitFit: Simple parameter-efficient fine-tuning for Transformer-based masked language-models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 1–9).

19. Mazumder, S., Liu, B., Ma, N., & Wang, H. (2019). Lifelong and interactive learning of factual knowledge in dialogues. In Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue (pp. 21–31).