Beyond the Window: Context Handling Mechanisms for Effective Human–Agent Collaboration

Lukas Röß
Munich University of Applied Sciences
Munich, Germany
lukas.roess@hm.edu

Abstract

Large language models increasingly act as partners in human agent collaboration. For such collaboration to be effective, agents must handle context over extended interactions in a way that supports shared understanding, reliable reasoning and appropriate trust. Recent work has explored several mechanisms for context handling, including sequencing through multi agent workflows, retrieval augmented generation, summarization and extended context windows. Yet it remains unclear how these mechanisms jointly influence the effectiveness of human agent collaboration rather than model accuracy in isolation.

This paper presents a focused literature synthesis of recent work from 2024 and 2025 on long context processing, memory and collaboration with large language models. We review multi agent sequencing approaches such as Chain of Agents, Graph of Agents and XpandA, which divide long inputs across agents and then recompose their contributions. We contrast these methods with empirical comparisons of retrieval augmented generation and long context models in industrial and clinical domains. We further integrate conceptual and empirical work on memory architectures and summarization, as well as studies of human teams supported by AI summarizers and investigations of real world human AI collaboration patterns.

Building on this synthesis, we propose a conceptual framework that relates four context handling mechanisms sequencing, retrieval, summarization and context window capacity to three dimensions of collaboration effectiveness: task performance, interaction quality and longitudinal alignment. We use this framework to refine the following research question for the remainder of the project. How do different mechanisms of context handling, such as sequencing, retrieval, summarization and context window capacity, influence the effectiveness of human agent collaboration? The paper concludes with a research design outline that will guide the next project phase.

CCS Concepts

• **Human-centered computing** → **Collaborative interaction**; • **Computing methodologies** → *Natural language processing*; • **Information systems** → *Information retrieval*.

Keywords

large language models, human–AI collaboration, context windows, retrieval augmented generation, summarization, memory

HM Seminar 2025, Munich, Germany
2025. ACM ISBN 000-0-0000-0000-0/25/00
https://doi.org/10.0000/placeholder-doi

ACM Reference Format:
Lukas Röß. 2025. Beyond the Window: Context Handling Mechanisms for Effective Human–Agent Collaboration. In *Proceedings of HM Seminar on Human–AI Collaboration (HM Seminar 2025).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.0000/placeholder-doi

1 Introduction

Large language models (LLMs) have moved from static tools toward agents that participate in ongoing collaboration with humans in domains such as analysis, writing and decision support. In these settings, human partners and LLM agents exchange information incrementally over many turns. The agent must preserve relevant context, recover earlier decisions and adapt to evolving goals rather than only solving one isolated prompt.

A central technical and design challenge is how LLM based agents handle context. Early work focused on extending the context window of individual models, yet this strategy can suffer from attention dilution and the well known tendency of models to under use information far from the end of the sequence. Recent multi agent approaches therefore explore sequencing and division of labour. Chain of Agents processes long inputs by passing segmented portions of the text through a chain of worker agents and then using a manager agent to integrate their contributions, which improves performance over retrieval based and full context baselines on long context tasks [9]. Graph of Agents generalizes this idea by constructing dynamic collaboration graphs that formalize long context processing as a compression problem and often outperform both retrieval augmented generation and fixed multi agent strategies while using much shorter local windows [1]. XpandA proposes a question driven multi agent framework that dynamically partitions long texts and coordinates agents through a central memory, enabling processing of sequences up to one million tokens while retaining competitive accuracy and latency [7].

In parallel, researchers compare retrieval augmented generation with long context models in realistic applications. An industry study shows that when resources are sufficient, long context models tend to outperform retrieval augmented generation across benchmarks, yet retrieval remains significantly more cost efficient and can be combined with long context models through routing strategies [3]. In the medical domain, a study on question answering demonstrates that carefully designed long context retrieval architectures can mitigate the usual loss of performance for information in the middle of a long sequence and improve answer quality over standard retrieval pipelines [8].

Beyond raw context capacity, there is growing interest in memory mechanisms and summarization. A recent survey synthesizes memory mechanisms for LLM agents including short term, long term and external memory, and emphasizes that memory must be

designed with the dynamics of human interaction in mind [10]. Conceptual work on contextual memory intelligence argues that memory should be treated as adaptive infrastructure that captures decisions and their context over time, supporting reflection, accountability and robust collaboration [6]. Empirical work from DeepMind explores how different forms of collaboration and memory banks interact with chain of thought reasoning and multi agent setups, showing that naive use of exemplars and memory can sometimes distract models rather than help them [4].

Studies of human teams interacting with AI systems complement these technical advances. A controlled experiment with two different AI summarizers for virtual collaborative analysis shows that summary style shapes trust and patterns of attention to teammate information even when final problem solving accuracy does not change [2]. A recent analysis of student interactions with LLMs on complex tasks reveals that collaboration often degenerates into instruct serve repeat sequences with long threads that show misalignment between human intent and model output rather than genuine negotiation or shared problem solving [5].

Taken together, these works indicate that context handling is not only a question of how many tokens a model can accept. Sequencing through multi agent workflows, retrieval strategies, summarization choices and the design of memory all interact with the human side of collaboration, including trust, attention and perceived agency. However, existing work tends to evaluate these mechanisms in isolation or on purely technical benchmarks. There is limited integrative analysis of how different context handling mechanisms jointly influence the effectiveness of human agent collaboration.

To address this gap, this project investigates the following research question.

Research question. How do different mechanisms of context handling, such as sequencing, retrieval, summarization and context window capacity, influence the effectiveness of human agent collaboration.

The contributions of this paper in the context of the seminar project are threefold. First, it summarizes recent work on long context processing, memory and collaboration into a structured view of context handling mechanisms. Second, it relates these mechanisms to dimensions of collaboration effectiveness that are relevant for human agent teaming. Third, it derives a concrete research design outline that will guide the empirical phase of the project in the next milestone.

2 Background

2.1 Human agent collaboration with LLMs

Human agent collaboration refers to settings in which humans and artificial agents jointly work on tasks with shared goals, roles and responsibilities. Modern LLM based agents can generate text, reason about documents, call tools and maintain conversational histories. This enables scenarios where the agent acts as partner rather than simple oracle.

However, recent empirical work suggests that current systems often fall short of this ideal. An analysis of student interaction sequences with LLM based tools in a complex problem solving task finds that most interactions follow an instruct serve repeat pattern. Students issue instructions, the AI responds and the process repeats, but there is limited negotiation, joint sensemaking or mutual adaptation [5]. This pattern indicates that context is not used to deepen the collaboration, even though long conversational histories are available.

In contrast, work on human teams supported by AI summarizers shows that context handling can substantially shape collaboration processes. In an experiment with virtual teams, two different AI summarizers produced either informative or indicative summaries of shared documents. While problem solving accuracy remained similar, the informative summaries increased trust in the AI and influenced how quickly and how often participants attended to new teammate information [2]. This result highlights that even simple choices about summarization and information presentation alter human attention and trust, which are central components of effective collaboration.

These findings motivate a closer examination of how context handling mechanisms in LLM agents interact with human behaviours and perceptions, rather than focusing only on model accuracy.

2.2 Context representation and memory in LLM agents

From a technical perspective, LLMs process context through four main strategies that are relevant for this project. The first strategy is to extend the context window so that more tokens can be provided to the model directly. Long context models like GPT or Gemini variants fall into this category. The second strategy is retrieval augmented generation, where external tools retrieve relevant documents or conversation snippets and inject them into a smaller context window. The third strategy is summarization, which compresses prior content into shorter representations that can be kept in the window or in external memory. The fourth strategy is explicit memory architectures that maintain information beyond a single prompt response cycle.

A comprehensive survey on memory mechanisms for LLM based agents distinguishes short term rolling context, episodic and semantic memory in vector stores, parametric memory via fine tuning and hybrid memory controllers. The survey emphasizes that memory systems must not only store content but also decide what to retain, when to update and how to retrieve information for different tasks and collaboration patterns [10]. This view supports the idea that sequencing, retrieval and summarization are not independent tricks but interacting components of a larger memory system.

Contextual memory intelligence extends this notion and argues that memory should be treated as a first class infrastructure capability for generative AI systems. It introduces a layered architecture with an insight layer that captures decisions, rationales and environmental signals over time, combined with human in the loop reflection and drift detection. The goal is to support longitudinal coherence, explainability and responsible collaboration in organizational settings [6]. This work explicitly situates memory design in human organizational contexts rather than narrow benchmarks.

Empirical work on collaboration and memory shows that the interaction between multi agent collaboration, chain of thought reasoning and memory banks can be nuanced. A DeepMind study compares fixed, random and similarity based retrieval of exemplars,

as well as frozen and continuously learned memory banks, in multi agent reasoning tasks. The results indicate that principled retrieval and memory are not always beneficial and that random exemplar selection can sometimes outperform more complex strategies. In some tasks, adding any exemplars distracts both weak and strong models [4]. These findings caution against assuming that more memory or more elaborate retrieval always improves performance or collaboration.

2.3 Mechanisms of context handling

For this project, we focus on four mechanisms of context handling that recur across the reviewed literature.

Sequencing. Sequencing refers to structuring processing as a sequence of steps, often across multiple agents with different roles. Chain of Agents introduces worker agents that each process segments of a long input, followed by a manager agent that synthesizes their outputs, which allows the system to exploit long documents while restricting the context for each individual agent [9]. Graph of Agents formalizes long context processing as a compression problem and constructs dynamic graphs where agents iteratively exchange and refine information according to an information theoretic objective [1]. XpandA uses a question driven workflow where agents iteratively query and update a shared memory about different partitions of an ultra long text, providing a form of sequenced collaboration across partitions [7].

Retrieval. Retrieval augmented generation retrieves relevant documents, conversation segments or exemplars from external stores and passes them to the model in a limited context window. The EMNLP industry study benchmarks retrieval augmented generation and long context models on multiple datasets and shows that long context models can outperform retrieval when computational resources are not constrained, yet retrieval can be more cost efficient and can be combined with long context through routing [3]. The medical question answering study extends this idea by combining long context with retrieval to mitigate the usual performance drop for information in the middle of the context and demonstrates improved answer quality [8].

Summarization. Summarization compresses content so that more information can be retained or shared within limited context budgets. In multi agent systems, summarization is used both for internal state and for communication between agents. Chain of Agents relies on the manager agent to summarize worker outputs into a final answer [9]. Graph of Agents uses compression objectives to guide how agents summarize and pass information along edges [1]. XpandA maintains summarized ensembles of information in a central memory and selectively replays partitions when needed [7]. In human teams, AI summarizers mediate what information humans see and how they allocate attention, as shown in the virtual collaborative analysis experiment [2].

Context window capacity. Context window capacity determines how many tokens can be processed at once. Extended context windows allow direct conditioning on long histories but do not guarantee that relevant information will be used effectively. Studies on long context models suggest that the performance benefits of larger windows depend on task structure and input distributions. The EMNLP industry study finds that long context models outperform retrieval only when resources are sufficient and that hybrid routing strategies are needed in practice [3]. The medical study shows that long context retrieval pipelines can leverage extended context to improve answer quality in complex clinical cases [8]. Multi agent approaches such as Graph of Agents show that carefully designed collaboration architectures can surpass much larger context windows by distributing processing across small local windows [1].

These four mechanisms form the conceptual basis for the analysis and research design in the remainder of the project.

3 Related work

3.1 Sequencing and multi agent collaboration

Multi agent approaches treat an LLM based system as a team of agents that each handle part of the context or reasoning process. Chain of Agents introduces a simple yet effective design where worker agents individually process segments of a long document and communicate their findings in sequence. A manager agent then aggregates these findings and produces a final answer. Experiments on long context question answering, summarization and code completion show that this approach can outperform retrieval augmented generation, full context operations and other multi agent baselines, even though each agent operates on a short local window [9]. This supports the idea that sequencing can compensate for limited context capacity.

Graph of Agents extends this design by constructing a dynamic graph of agents whose communication pattern is guided by an information theoretic compression objective. Rather than following a fixed chain, agents are connected according to the structure of the input and iteratively refine compressed representations. On long context document question answering benchmarks, this approach improves F1 scores over both retrieval augmented generation and a strong fixed multi agent baseline, while effectively increasing the usable context beyond the base model limit [1]. The work highlights that emergent collaboration structures can be more effective than hand crafted pipelines.

XpandA focuses on ultra long sequences up to one million tokens. It partitions texts dynamically so that each partition fills the context window to a controlled degree and uses a question driven workflow where agents update a flat ensemble of information in a shared memory. Agents selectively replay partitions when necessary to resolve temporal structures such as flashbacks. This design achieves substantial improvements over full context models, retrieval augmented generation and previous agent based methods in both accuracy and inference speed [7]. XpandA therefore shows how sequencing, shared memory and dynamic partitioning jointly scale long context processing.

3.2 Retrieval augmented generation and long context models

The trade off between retrieval augmented generation and extended context windows is central for context handling. The EMNLP industry study conducts a large scale comparison of these strategies using several contemporary LLMs across diverse datasets. The authors report that when model and hardware resources are sufficiently provisioned, long context models consistently outperform retrieval augmented generation in terms of average performance. However,

retrieval remains attractive due to significantly lower cost. The study proposes a hybrid routing method that sends queries either to retrieval augmented pipelines or to long context inference based on model self reflection, achieving near long context performance at reduced cost [3]. This work is important for this project because it quantifies the trade offs between context capacity and retrieval under realistic constraints.

In the clinical domain, long context retrieval architectures are studied for medical question answering. The npj Digital Medicine paper combines retrieval augmented language models with extended context windows and designs a method to counter the typical degradation of performance for information located in the middle of long contexts. By leveraging long context representations in combination with retrieval, the authors improve answer quality on challenging medical datasets compared to standard retrieval augmented generation baselines [8]. This result demonstrates that context window capacity and retrieval can be complementary rather than mutually exclusive mechanisms.

These works suggest that effective human agent collaboration in high stakes domains will likely rely on hybrid systems that decide when to use retrieval, when to rely on long context and how to combine both within an interaction.

3.3 Memory architectures and summarization

Memory architectures aim to preserve information across interactions. The survey on memory mechanisms for LLM based agents synthesizes work on sliding window context, episodic memory in vector stores, semantic memory graphs, learned key value caches and other structures. It organizes these mechanisms by retention horizon and access pattern, and emphasizes the tension between storing rich context and avoiding distraction or hallucination from outdated content [10]. This survey provides the conceptual foundation for treating sequencing, retrieval, summarization and window size as components of a memory system rather than isolated tricks.

Contextual memory intelligence introduces a more normative perspective. It argues that memory should be designed to support longitudinal coherence, explanation and governance in organizational settings. The proposed insight layer captures decisions, rationales and environment changes, and integrates human in the loop review and drift detection. The paper positions memory as a key enabler for human agent collaboration that remains aligned with institutional constraints over time [6]. This view motivates the inclusion of collaboration effectiveness metrics beyond accuracy in the present project.

Empirical work on collaboration and memory from DeepMind connects these ideas to concrete performance trade offs. In controlled experiments, the authors vary collaboration styles between agents, the design of memory banks and retrieval methods, and measure reasoning performance on grounded tasks. Results show that chain of thought reasoning, multi agent collaboration and memory banks interact in complex ways. Principled similarity based retrieval does not always outperform random exemplar selection, and in certain tasks memory can even distract models [4]. These findings emphasize that memory and summarization mechanisms must be evaluated jointly with the tasks and collaboration patterns they support.

3.4 Human teams with AI support

The small group research literature provides insights into how AI mediated context handling affects human teams. In a study of virtual collaborative analysis, teams used a platform where an AI summarizer presented either an informative summary that synthesized key content or an indicative summary that provided a lighter overview. The experiment found that participants with informative summaries developed higher trust in the AI and changed their attention patterns to teammate information, although problem solving accuracy remained similar between conditions [2]. This indicates that summarization style influences social dynamics and perceptions even when task level outcomes remain stable.

Another study examines how students interact with LLM systems while solving a complex task. Using sequence analysis and network techniques, the authors show that interaction patterns are dominated by simple instruct serve repeat dynamics. Long message threads often reveal persistent misalignment between prompts and AI responses, and the analysis finds no strong relationship between assignment complexity, prompt length and student performance [5]. The study concludes that current LLM systems, optimized for following instructions, are not yet strong partners for cognitive collaboration.

These results underline that context handling mechanisms must be studied not only as algorithmic design choices but also as factors that shape trust, attention and perceived agency in human agent collaboration.

4 Conceptual framework

Based on the reviewed literature, we propose a conceptual framework that relates context handling mechanisms to collaboration effectiveness.

We consider four mechanisms: sequencing, retrieval, summarization and context window capacity. We examine their influence along three dimensions of collaboration effectiveness.

Task performance. This includes accuracy, completeness and efficiency on joint tasks. Multi agent sequencing improves performance on long context question answering and summarization compared to retrieval augmented baselines [1, 7, 9]. Long context retrieval architectures enhance answer quality in medical question answering [8]. Hybrid routing between retrieval and long context can approach the performance of pure long context systems at lower cost [3].

Interaction quality. This dimension captures trust, attention and communication patterns in human agent teaming. Informative AI summaries can increase trust and alter attention to teammate information [2]. In contrast, real world interactions with LLMs often show limited negotiation and misalignment between human intent and AI output [5]. Memory and summarization mechanisms can either support or hinder coordinated reasoning in multi agent systems [4].

Longitudinal alignment. This dimension concerns how well collaboration remains coherent and aligned over longer periods. Memory surveys and contextual memory intelligence argue that memory mechanisms should support stable yet adaptable representations of context, decisions and rationales [6, 10]. Multi agent frameworks like XpandA demonstrate how shared memory and

partition replay can preserve consistency across very long inputs [7].

The refined research question asks how different combinations of the four mechanisms influence these three dimensions of effectiveness in human agent collaboration scenarios.

5 Research design outline

This section provides an outline of the planned empirical work for the next project phase. The details will be refined in the Milestone 3 plan, but the structure is defined here to complete the Milestone 2 deliverable.

5.1 Study design

The planned study will compare four conditions corresponding to different context handling mechanisms in a controlled human agent collaboration task, for example collaborative document analysis or scenario planning.

- **Long context only.** A single long context LLM agent operates with an extended context window but no external retrieval or explicit memory.
- **Retrieval augmented generation.** A smaller context LLM agent uses retrieval from a document store and conversation history to answer questions and co author content.
- **Summarization focused agent.** An agent that aggressively summarizes prior context into short notes and exposes these summaries to the human partner, inspired by findings on AI summarizers in team settings [2].
- **Sequenced multi agent system.** A simplified Chain of Agents or XpandA inspired pipeline where several agents read different parts of the input and a coordinator agent synthesizes their contributions, with a shared memory component [7, 9].

Participants will work with one of these systems on a multi step task that requires integrating information from multiple documents and adapting to evolving instructions.

5.2 Measures

To align with the conceptual framework, the study will collect measures for the three dimensions of effectiveness.

- **Task performance.** Objective metrics such as answer accuracy, coverage of relevant information, time to completion and number of interaction turns, following practices in long context benchmarks [1, 3, 9].
- **Interaction quality.** Self report scales for trust in the agent, perceived understanding and workload, as well as behavioural measures such as distribution of attention across sources and types of interaction patterns, informed by the AI summarizer study and the instruct serve repeat analysis [2, 5].
- **Longitudinal alignment.** Measures of consistency across task phases, such as whether the agent maintains prior decisions and rationales, inspired by contextual memory frameworks [6, 10].

5.3 Analysis plan

The analysis will compare conditions along the three dimensions using appropriate statistical tests, for example mixed effects models for repeated measures. It will also examine interaction patterns qualitatively and through sequence analysis to detect collaboration modes similar to those described in prior work [5]. The results will be interpreted using the proposed framework to understand how combinations of sequencing, retrieval, summarization and context window capacity shape human agent collaboration.

6 Conclusion and outlook

This paper updated the project focus in line with feedback that context management and sequencing are at least as important as raw context window size for human agent collaboration. Through a targeted literature review of recent work from 2024 and 2025, we synthesized contributions on multi agent sequencing for long context processing, empirical comparisons of retrieval augmented generation and long context models, memory and summarization mechanisms, and empirical studies of human teams interacting with AI systems.

The reviewed work suggests that effective collaboration depends on how context is structured, retrieved, summarized and preserved over time, not only on the number of tokens an LLM can process at once. Multi agent frameworks such as Chain of Agents, Graph of Agents and XpandA demonstrate that sequencing and shared memory can extend effective context beyond model limits. Industrial and medical studies show that hybrid combinations of retrieval and long context are promising for real applications. Memory surveys and contextual memory intelligence frameworks argue that memory should be treated as adaptive infrastructure for responsible collaboration. Human team studies highlight that summarization style and interaction patterns strongly influence trust and perceived collaboration quality.

Building on this synthesis, we formulated a research question that explicitly targets the joint influence of sequencing, retrieval, summarization and context window capacity on the effectiveness of human agent collaboration. We proposed a conceptual framework and outlined a comparative study that will operationalize this question through controlled human agent experiments.

In the next project phase, we will refine the experimental design, implement the agent variants and pilot the measures. The final goal is to provide actionable guidance on how to combine context handling mechanisms in LLM based systems that function as genuinely collaborative partners rather than obedient yet often misaligned assistants.

Acknowledgments

I would like to thank Prof. Dr. Hanna Moser and Prof. Dr. Gudrun Socher for their guidance in refining the research question. I used an AI assistant (ChatGPT, GPT-5.1 Thinking) to support language polishing, structuring of the outline, and drafting section text. All conceptual decisions, interpretation of the literature, and the final research design are my own.

This work was conducted as part of the seminar on human agent collaboration at Munich University of Applied Sciences.

References

[1] Taejong Joo, Shu Ishida, Ivan Sosnovik, Bryan Lim, Sahand Rezaei-Shoshtari, Adam Gaier, and Robert Giaquinto. 2025. Graph of Agents: Principled Long Context Modeling by Emergent Multi-Agent Collaboration. *arXiv preprint* (2025). arXiv:2509.21848 [cs.CL] https://arxiv.org/abs/2509.21848

[2] Aimée A. Kane, Susannah B. F. Paletz, Madeline Diep, Alexander Hajkowski, and Adam A. Porter. 2025. Virtual Collaborative Analysis: Effects of Two AI Summarizers. *Small Group Research* (2025). doi:10.1177/10464964251361563 Advance online publication.

[3] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Miami, Florida, USA, 881–893. https://aclanthology.org/2024.emnlp-industry.66

[4] Julie Michelman, Nasrin Baratalipour, and Matthew Abueg. 2025. Enhancing Reasoning with Collaboration and Memory. *arXiv preprint* (2025). arXiv:2503.05944 [cs.CL] https://arxiv.org/abs/2503.05944

[5] Mohammed Saqr, Kamila Misiejuk, and Sonsoles López-Pernas. 2025. Human-AI Collaboration or Obedient and Often Clueless AI in Instruct, Serve, Repeat Dynamics? *arXiv preprint* (2025). arXiv:2508.10919 [cs.HC] https://arxiv.org/abs/2508.10919

[6] Kristy Wedel. 2025. Contextual Memory Intelligence: A Foundational Paradigm for Human-AI Collaboration and Reflective Generative AI Systems. *arXiv preprint* (2025). arXiv:2506.05370 [cs.AI] https://arxiv.org/abs/2506.05370

[7] Sibo Xiao, Zixin Lin, Wenyang Gao, and Yue Zhang. 2025. Long Context Scaling: Divide and Conquer via Multi-Agent Question-driven Collaboration. *arXiv preprint* (2025). arXiv:2505.20625 [cs.CL] https://arxiv.org/abs/2505.20625

[8] Gongbo Zhang, Zihan Xu, Qiao Jin, Fangyi Chen, Yilu Fang, Yi Liu, Justin F. Rousseau, Ziyang Xu, Zhiyong Lu, Chunhua Weng, and Yifan Peng. 2025. Leveraging Long Context in Retrieval Augmented Language Models for Medical Question Answering. *npj Digital Medicine* 8, 239 (2025). doi:10.1038/s41746-025-01651-w

[9] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö. Arik. 2024. Chain of Agents: Large Language Models Collaborating on Long-Context Tasks. In *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2024/file/ee71a4b14ec26710b39ee6be113d7750-Paper-Conference.pdf NeurIPS 2024.

[10] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A Survey on the Memory Mechanism of Large Language Model-based Agents. *ACM Transactions on Information Systems* 43, 6 (2025), 1–45. doi:10.1145/3748302