# Advances in In-Context Learning for LLMs (2025–Present)

## Introduction

Large Language Models (LLMs) have demonstrated an *in-context learning* ability – they can adapt to new tasks or incorporate new information at inference time by conditioning on prompts, without any gradient updates to their weights[[1]](https://arxiv.org/abs/2507.16003#:~:text=,reason%20why%20LLMs%20can%20learn)[[2]](https://icml.cc/virtual/2025/poster/44826#:~:text=Large%20language%20models%20like%20ChatGPT,depending%20on%20the%20question%20they%E2%80%99re). This dynamic learning within the context window allows models to “learn” from examples or new facts provided in the prompt, mimicking training-time learning but on the fly. Recent research (2025 and beyond) has focused on **enhancing in-context learning** so that LLMs can robustly integrate new information and skills *within their context window*, rather than relying on expensive retraining or fine-tuning. Key directions include: designing **synthetic or persistent memory** mechanisms to carry knowledge across queries, understanding the **implicit adaptation** happening inside transformer layers, improving **retrieval-based methods** that feed external knowledge into the prompt, and enabling **continual in-context learning** across extended interactions. This report reviews the latest advances in architectures, mechanisms, benchmarks, and theoretical foundations that empower LLMs to dynamically acquire and use new information via context alone, without updating their pretrained weights.

## Synthetic Memory and Persistent In-Context Learning

A major 2025 trend is augmenting LLMs with **persistent memory** that lives outside the model weights but within or alongside the context. Rather than treating each query in isolation, new frameworks allow an LLM to accumulate and reuse knowledge over multiple queries. **Dynamic Cheatsheet (DC)** by Suzgun *et al.* (2025) is one such framework that gives a black-box LLM an evolving textual memory at inference time[[3]](https://arxiv.org/html/2504.07952v1#:~:text=Despite%20their%20impressive%20performance%20on,truth%20labels%20or%20human)[[4]](https://arxiv.org/html/2504.07952v1#:~:text=Engineering%20and%20Physics%20problems,driven%20learning%20characteristic%20of%20human). After solving each query, the model “curates” helpful intermediate results (e.g. reasoning steps, code snippets) into a compact memory store. On subsequent queries, the model retrieves relevant snippets from this store to avoid re-deriving the same solutions[[5]](https://arxiv.org/html/2504.07952v1#:~:text=without%20retaining%20insights%20from%20previous,retaining%20algebraic%20insights%20across%20questions)[[6]](https://arxiv.org/html/2504.07952v1#:~:text=challenges%2C%20DC%20yields%20notable%20accuracy,responses%20and%20cutting%20routine%20errors). This test-time learning approach led to dramatic gains: for instance, feeding back correct Python code for arithmetic puzzles boosted GPT-4’s success on a math game from 10% to 99% by allowing it to reuse the discovered solution instead of starting from scratch[[7]](https://arxiv.org/html/2504.07952v1#:~:text=execution%2C%20but%20lacks%20a%20memory,memory%20augmentation%20and%20efficient%20retrieval)[[8]](https://arxiv.org/html/2504.07952v1#:~:text=An%20even%20more%20striking%20and,manually%20fumbling%20with%20numeric%20manipulations). Crucially, the DC memory is *self-curated* and succinct (storing only transferable “lessons” rather than entire dialogues), which avoids context overflow while facilitating meta-learning from past successes and failures[[4]](https://arxiv.org/html/2504.07952v1#:~:text=Engineering%20and%20Physics%20problems,driven%20learning%20characteristic%20of%20human)[[9]](https://arxiv.org/html/2504.07952v1#:~:text=We%20also%20note%20that%20DC,Roediger%20%26%20Butler%2C%202011%3B%20Karpicke). Because no model parameters are updated – the LLM simply consults and revises an external memory between queries – this approach **“adapts [the LLM’s] skills on the fly, without modifying underlying parameters,”** continuously improving performance over time[[10]](https://arxiv.org/html/2504.07952v1#:~:text=thereby%20facilitating%20meta,data%2C%20results%2C%20and%20code%20at). Such synthetic memory systems bridge the gap between isolated prompt-response pairs and the cumulative learning humans exhibit by building on experience[[10]](https://arxiv.org/html/2504.07952v1#:~:text=thereby%20facilitating%20meta,data%2C%20results%2C%20and%20code%20at).

Several other works integrate long-term memory into LLM-driven agents to enable **continual in-context learning**. *MemoryBank* (AAAI 2024) introduced a dedicated memory module for storing conversation history and important facts, which the agent can retrieve and even **update continuously** as interactions proceed[[11]](https://arxiv.org/html/2504.15965#:~:text=limitation%20of%20LLM%E2%80%99s%20context%20window%2C,61)[[12]](https://arxiv.org/html/2504.15965#:~:text=task%20adaptability%20and%20contextual%20awareness,like). This allows an agent to form a persistent user-specific profile and remember past events, overcoming the context window limit. MemoryBank uses a dual-tower retriever (FAISS-indexed vectors) to efficiently find relevant past information[[13]](https://arxiv.org/html/2504.15965#:~:text=match%20at%20L759%20MemoryBank%C2%A0,based%20retrieval), and even implements a decay/update schedule inspired by human forgetting curves to manage which memories to reinforce or discard over time[[12]](https://arxiv.org/html/2504.15965#:~:text=task%20adaptability%20and%20contextual%20awareness,like). Similarly, frameworks like **memoized chatbots** (e.g. OpenAI’s *ChatGPT with long-term memory*, Apple’s *Personal Context*, and open-source projects like *MemoryScope* and *mem0*) emerged to keep track of user-specific information beyond a single session[[14]](https://arxiv.org/html/2504.15965#:~:text=limitation%20of%20LLM%E2%80%99s%20context%20window%2C,21%5D%2C%20etc). These systems typically store dialogue summaries or key facts externally and retrieve them into the prompt when relevant, allowing the LLM to maintain continuity across sessions[[15]](https://arxiv.org/html/2504.15965#:~:text=I%20%20%20Personal%20Non,42)[[16]](https://arxiv.org/html/2504.15965#:~:text=Retrieval%20%20%20RET,124). By treating the context window as a short-term working memory and leveraging external stores as long-term memory, LLM agents can learn **continuously** from interactions without weight updates[[17]](https://arxiv.org/html/2504.15965#:~:text=executing%20tools%2C%20thereby%20efficiently%20completing,addition%2C%20many%20commercial%20and%20open)[[18]](https://arxiv.org/abs/2501.07278#:~:text=core%20components%20of%20these%20agents,roadmap%20for%20researchers%20and%20practitioners). A recent survey by Zheng *et al.* (2025) on *Lifelong Learning for LLM-based Agents* emphasizes that a **memory module for storing and retrieving evolving knowledge** is one of three core pillars (alongside perception and action) required to enable continuous adaptation and mitigate catastrophic forgetting in LLM agents[[19]](https://arxiv.org/abs/2501.07278#:~:text=This%20survey%20is%20the%20first,roadmap%20for%20researchers%20and%20practitioners). As LLM agents accumulate experiences, memory-augmented prompting ensures they recall past lessons and *incrementally* improve, rather than resetting at each prompt.

However, these memory systems face limitations. If the model is **too small or not sufficiently capable**, feeding it a memory of past errors or strategies may not help – Suzgun *et al.* note that smaller models often populate memory with flawed content and fail to correct it[[20]](https://arxiv.org/html/2504.07952v1#:~:text=approach%20instead%20of%20manually%20fumbling,with%20numeric%20manipulations). There is also the challenge of **memory management**: deciding what to store or forget to prevent an ever-growing context. Approaches like Dynamic Cheatsheet address this via selective curation (storing only concise, reusable insights)[[4]](https://arxiv.org/html/2504.07952v1#:~:text=Engineering%20and%20Physics%20problems,driven%20learning%20characteristic%20of%20human)[[9]](https://arxiv.org/html/2504.07952v1#:~:text=We%20also%20note%20that%20DC,Roediger%20%26%20Butler%2C%202011%3B%20Karpicke). Ensuring the memory remains accurate and not misleading is another concern, as **erroneous learned snippets could degrade performance** if reused without caution. Despite these challenges, synthetic memory approaches are a promising path to give LLMs a form of “experience” – enabling learning that persists across queries without any gradient-based updates.

## Implicit Adaptation Mechanisms in Transformers

While external memory provides one route to dynamic learning, another line of research asks: *How do LLMs inherently achieve in-context learning with frozen weights?* Recent theoretical works (2025) have started to demystify the **implicit adaptation** happening inside transformer models during inference. A landmark study by Dherin *et al.* (2025) showed that a transformer block can internally simulate a gradient descent step on itself – effectively performing a *hidden weight update* in response to the prompt[[21]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=Recent%20work%20by%20Dherin%20et,Intuitively%2C%20we%20can%20imagine%20that)[[22]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=every%20prompt%20example%20attaches%20a,instructions%20that%20alter%20its%20computations). Specifically, they proved that the combination of self-attention followed by an MLP in a transformer can *implicitly modify* the MLP’s effective weights based on the input context, without changing the stored parameters[[23]](https://arxiv.org/abs/2507.16003#:~:text=learn%20new%20patterns%20without%20any,Specifically%2C%20we%20show%20under%20mild). Under mild assumptions, each new token’s information induces a low-rank “update” to the MLP weight matrix within the forward pass[[24]](https://arxiv.org/abs/2507.16003#:~:text=seen%20during%20training,update%20of%20the%20MLP%20layer)[[21]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=Recent%20work%20by%20Dherin%20et,Intuitively%2C%20we%20can%20imagine%20that). We can intuitively think of each prompt example attaching a temporary **“patch” onto the model’s weights** (like a sticky note) that biases the model’s computation toward the pattern in that example[[25]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=viewed%20as%20performing%20a%20form,disappear%20and%20the%20base%20model). Once inference is done, these implicit weight tweaks disappear – the base model remains unchanged – but during processing they *bridge the gap between static weights and dynamic behavior*, allowing the model to effectively **“learn without training” by using the prompt as instructions to alter its internal computation**[[22]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=every%20prompt%20example%20attaches%20a,instructions%20that%20alter%20its%20computations). This insight provides a theoretical foundation for in-context learning: it is not just pattern matching or retrieval of memorized examples, but a genuine on-the-fly *optimization*. It explains how an LLM can generalize to prompt demonstrations it never saw in training – the transformer is performing an *inner-loop learning* step via its forward pass[[26]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=prompt%20examples%20as%20if%20it,For%20example%2C%20even%20when%20prompt)[[27]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=The%20work%20of%20Dherin%20et,focus%20on%20how%20the%20self).

Complementing this view, Wu *et al.* (ICML 2025) interpret in-context learning through the lens of **associative memory**. They show that providing prompt examples is analogous to “*reshaping*” the energy landscape of a Hopfield network (a classic associative memory model) to bias retrieval towards those examples[[28]](https://icml.cc/virtual/2025/poster/44826#:~:text=Abstract%3A%20We%20provide%20an%20exactly,context%20examples%20shape). In their analysis, each in-context example subtly shifts what the model “remembers,” focusing it on patterns relevant to the query[[29]](https://icml.cc/virtual/2025/poster/44826#:~:text=%E2%80%9Clearn%E2%80%9D%20so%20quickly%20without%20updating,context%20learning%20is)[[30]](https://icml.cc/virtual/2025/poster/44826#:~:text=process%20can%20be%20understood%20as,depending%20on%20the%20question%20they%E2%80%99re). In other words, the transformer’s attention mechanism can be seen as retrieving from an implicit memory of training examples, but *conditioned* on the prompt so that memories congruent with the prompt are activated[[31]](https://icml.cc/virtual/2025/poster/44826#:~:text=process%20can%20be%20understood%20as,drawn%20from%20a%20familiar%20setting). This perspective helps explain phenomena like why in-context learning works better when prompt examples are similar to the test query and accurate – because those factors make the induced memory retrieval more effective[[32]](https://icml.cc/virtual/2025/poster/44826#:~:text=like%20how%20a%20person%20recalls,drawn%20from%20a%20familiar%20setting). Together, these theoretical advancements (Bayesian interpretations, implicit SGD in transformers, and associative memory analogies) deepen our understanding of *how* LLMs learn from context and guide new improvements.

Researchers have also begun modifying model **architectures** to enhance this implicit adaptation. Burns *et al.* (2025) drew inspiration from biological memory systems to design a *residual attention stream* that allows information to flow directly between attention heads[[33]](https://openreview.net/forum?id=lcTFm4LIRR#:~:text=computational%20neuroscience%20community%20to%20model,larger%20and%20more%20naturalistic%20scales). By connecting attention outputs in a novel way, a two-layer transformer with this architecture learned in-context tasks faster and more effectively than a standard transformer[[34]](https://openreview.net/forum?id=lcTFm4LIRR#:~:text=ICL,values%2C%20with%20results%20also%20indicating)[[35]](https://openreview.net/forum?id=lcTFm4LIRR#:~:text=LLMs,8%20million%20and%201%20billion). In small LMs (8M–1B parameters), their modifications led to earlier and stronger emergence of in-context learning capabilities[[36]](https://openreview.net/forum?id=lcTFm4LIRR#:~:text=ICL,larger%20and%20more%20naturalistic%20scales). This suggests architectural tweaks can lower the scale threshold for ICL or make models more efficient in adapting to prompts. Another avenue has been extending context length via efficient attention or recurrence, as seen in models like **Retentive Networks (RetNet)**. RetNet (initially proposed in late 2023) replaces self-attention with a retention mechanism that has linear complexity and a form of statefulness, enabling context windows of hundreds of thousands of tokens[[37]](https://research.ibm.com/blog/larger-context-window#:~:text=Why%20larger%20LLM%20context%20windows,more%20coherent%20and%20relevant%20answers). Longer contexts directly translate to more “working memory” for in-context learning – indeed, an LLM with a 100K token window can ingest entire knowledge bases or lengthy demonstrations at test time. By 2025, several LLMs (e.g. *Qwen-Long*, *GPT-4 128K*) began offering massive context lengths, allowing them to *dynamically* integrate far more information without weight updates[[38]](https://magazine.sebastianraschka.com/p/llm-research-papers-2025-list-one#:~:text=%2A%2023%20May%2C%20QwenLong,17667). The trade-off is computational: extremely long contexts can slow inference, and the model must be trained or adapted to utilize them effectively (avoiding dilution of attention). Nonetheless, architectural innovations for long-context processing work hand-in-hand with in-context learning methods – together enabling models to take in and reason over new information of substantial scope, on the fly.

## Retrieval-Augmented and External Knowledge Approaches

**Retrieval-Augmented Generation (RAG)** – feeding an LLM with relevant external documents – is an established strategy (pre-2025) to inject up-to-date knowledge without retraining. Recent advances seek to make retrieval **more dynamic, contextual, and integrated with reasoning**. Traditional RAG pipelines retrieve text purely based on the query, and while they expand factual coverage, they do not teach the model *how to use* those facts in reasoning[[39]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=Existing%20fixes%20have%20their%20own,on%20how%20to%20use%20it). Two notable 2025 developments address this gap:

* **RARE (Retrieval-Augmented Reasoning Enhancement)**[[40]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=A%20team%20of%20researchers%20from,it%20down%20in%20simple%20terms)[[41]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=Enter%20RARE%3A%20A%20New%20Way,Forward): This approach “flips the script” of RAG by explicitly splitting the problem into *knowledge retrieval* and *reasoning*, and focusing the model’s training on the *latter*. Wang *et al.* (2025) in RARE propose to **outsource knowledge storage** to an external database while training the LLM (even a relatively small one, ~7B params) intensively on *how to reason with retrieved evidence*[[41]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=Enter%20RARE%3A%20A%20New%20Way,Forward)[[42]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=outsources%20knowledge%20storage%20and%20focuses,level%20thinking%20over). During training, the model is fed queries alongside relevant documents and is supervised to produce step-by-step reasoning (chains-of-thought) that arrive at correct answers[[43]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=,bogged%20down%20memorizing%20facts%2C%20it)[[44]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=1.%20High,world%20challenges). The result is a lean model specialized in logical problem-solving, paired with a plug-in knowledge source for facts. This separation of concerns yielded striking results: a RARE-trained 7B model (Qwen-7B variant) **outperformed** a retrieval-augmented GPT-4 on medical QA benchmarks (78.6% vs 75.2% on PubMedQA)[[45]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=%2A%20Better%20Performance%3A%20A%20RARE,without%20retraining%20the%20whole%20AI). Despite being orders of magnitude smaller, the RARE model’s focused reasoning training allowed it to make better use of retrieved data, solving domain-specific questions more accurately than a generic LLM[[45]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=%2A%20Better%20Performance%3A%20A%20RARE,without%20retraining%20the%20whole%20AI). RARE models were also more *efficient* (no need to cram every fact into parameters) and *flexible*, since updating the external knowledge base (e.g. adding new medical research) immediately updates the system’s knowledge without any model retraining[[46]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=That%E2%80%99s%20a%20smaller%20AI%20outperforming,without%20retraining%20the%20whole%20AI). In essence, RARE is a **“paradigm shift where maintainable external knowledge bases synergize with compact, reasoning-optimized models.”**[[47]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=The%20RARE%20framework%20shows%20that,to%20make%20AI%20an%20expert) It addresses both knowledge completeness and reasoning robustness: the model is no longer a static encyclopedia, but an adaptable reasoner that can plug in new info as needed.
* **Dynamic Contextual Retrieval & Tool Use**: Another thread has improved how retrieval is done in multi-turn or tool-augmented settings. *Dynamic Context Tuning (DCT)* (Arxiv Jun 2025) is a framework for dialogue-based assistants that integrates retrieval with memory and tool usage. DCT introduced a **“multi-turn context cache”** to track dialogue history and resolve coreferences, a **domain-adaptive retriever** (using LoRA fine-tuning) to quickly incorporate new tools or APIs, and **lightweight context compression** to keep prompts concise[[48]](https://arxiv.org/html/2506.11092v1#:~:text=To%20address%20these%20issues%2C%20we,DCT%20introduces%20three%20key%20innovations)[[49]](https://arxiv.org/html/2506.11092v1#:~:text=Multi,turn%20understanding). This dynamic retrieval system lets an assistant maintain context over long conversations and even adapt to previously unseen tools without full retraining[[50]](https://arxiv.org/html/2506.11092v1#:~:text=Moreover%2C%20DCT%20generalizes%20to%20previously,wide%20range%20of%20dynamic%20environments)[[51]](https://arxiv.org/html/2506.11092v1#:~:text=On%20the%20tool%20adaptation%20front%2C,without%20retraining%20the%20base%20model). In evaluations on support tasks, DCT improved multi-turn accuracy by 14% and cut hallucinations by over a third, while being able to plug in new APIs on the fly[[52]](https://arxiv.org/html/2506.11092v1#:~:text=We%20evaluate%20DCT%20on%20tasks,constrained%20settings)[[50]](https://arxiv.org/html/2506.11092v1#:~:text=Moreover%2C%20DCT%20generalizes%20to%20previously,wide%20range%20of%20dynamic%20environments). The key innovation is treating retrieved knowledge (or tool results) as part of an evolving dialog state that the model **learns to manage** (through context caching and adaptive retrieval) rather than a static one-off lookup. This blurs the line between pure retrieval and “learning from the retrieved info” within the conversation. Similarly, retrieval strategies like *Self-RAG* and *FLARE* were extended in 2024–2025 to plan multi-hop queries or anticipate future information needs during a conversation[[53]](https://arxiv.org/html/2506.11092v1#:~:text=Conversational%20RAG%20systems%20extend%20retrieval,not%20support%20active%20tool%20invocation). All these advances aim to make retrieval more **context-aware and continuous**, enabling the model to incorporate new information fluidly as a dialogue or task progresses. The combination of retrieval with chain-of-thought prompting has proven especially powerful in domains like troubleshooting, where the model can fetch specific facts and then reason step-by-step using them.

Despite the success of retrieval-based methods in injecting new knowledge, they do come with limitations. One is the dependence on a quality retriever and knowledge source – poor retrieval will feed irrelevant or incorrect info, which the LLM may then faithfully incorporate (garbage in, garbage out). Thus, research into better retrievers (e.g. learning retriever parameters with feedback from the LLM’s performance) is ongoing[[51]](https://arxiv.org/html/2506.11092v1#:~:text=On%20the%20tool%20adaptation%20front%2C,without%20retraining%20the%20base%20model). Another challenge is that while retrieval provides facts, **it doesn’t inherently give the model new skills**. The RARE approach addresses this by explicitly training reasoning with retrieved data, but traditional RAG won’t help an LLM perform a completely novel task unless examples or instructions for that task are also retrieved. In-context learning of *skills* (procedures, algorithms) may require demonstration data, not just factual retrieval. Finally, large-scale retrieval introduces latency and complexity to deployments. Techniques like context compression and caching in DCT[[54]](https://arxiv.org/html/2506.11092v1#:~:text=Domain,8)[[49]](https://arxiv.org/html/2506.11092v1#:~:text=Multi,turn%20understanding) help mitigate the overhead by summarizing or selectively retaining past info. Going forward, the integration of retrieval with other test-time learning mechanisms (memory, self-refinement, etc.) is an exciting frontier – systems that can not only fetch facts but also **critique and update their own outputs** using external knowledge (a sort of self-retrieval and revision loop).

## Continual and Lifelong In-Context Learning

An ultimate goal is for LLM-based agents to **learn continually from experience** – to improve over time as they handle more queries or interact with users, all without ever pausing for weight retraining. The developments discussed above (persistent memory, dynamic retrieval, etc.) are building blocks toward this vision. In 2025, we also see the emergence of benchmarks and frameworks explicitly targeting *lifelong learning* in LLMs. **LifelongAgentBench** (2025) is one example: a benchmark designed to evaluate how well an LLM agent can accumulate knowledge and skills over a sequence of tasks or dialogues, testing for issues like catastrophic forgetting and the stability-plasticity tradeoff (retaining old info vs. learning new)[[55]](https://www.nb-data.com/p/advancing-ai-lifelong-learning-roadmap#:~:text=1,knowledge%20when%20acquiring%20new%20information)[[56]](https://www.nb-data.com/p/advancing-ai-lifelong-learning-roadmap#:~:text=In%20contrast%20to%20traditional%20LLMs%2C,changing%20scenarios). The formulation of such benchmarks has been driven by the recognition that existing LLMs, while powerful, are usually *static* – they solve tasks with fixed knowledge and cannot integrate new information on the fly[[57]](https://www.nb-data.com/p/advancing-ai-lifelong-learning-roadmap#:~:text=Large%20Language%20Models%20,their%20effectiveness%20in%20dynamic%20settings). LifelongAgentBench scenarios force the model to incorporate new facts introduced mid-conversation or adapt to evolving goals, measuring whether it can do so within its context memory. Initial results often show that vanilla LLMs struggle, underscoring the need for the kind of methods reviewed in this report.

The *roadmap for lifelong LLM agents* by Zheng *et al.* (2025) encapsulates the research directions being pursued: it highlights enhanced perception (multimodal inputs), dedicated memory (for accumulating experiences), and action interfaces (tool use, environment interaction) as key components to achieve continual learning[[19]](https://arxiv.org/abs/2501.07278#:~:text=This%20survey%20is%20the%20first,roadmap%20for%20researchers%20and%20practitioners). Notably, many lifelong learning approaches for LLMs rely on the *synergy* of those components. For example, an agent might perceive a new piece of information (say, via a document or user message), store it in long-term memory, and later retrieve it to inform its actions – thereby “learning” a fact permanently without weight updates. Some experimental agent systems even implement **automated self-reflection or self-correction loops**: the LLM, after producing an answer, can re-read the conversation (or its own solution) and refine it, effectively learning from its mistakes in context. This kind of *implicit self-training* at inference time, often via chain-of-thought and feedback, has been used to boost performance on reasoning tasks. It connects to the idea of *synthetic experience*: the model generates and then learns from its own intermediate outputs, all within the context window.

Despite progress, fully achieving lifelong learning for LLMs faces open challenges. **Context length limitations** remain a fundamental bottleneck – even a 100K token context will eventually be exhausted if an agent never forgets anything. Strategies for **memory management and compression** (summarizing older interactions, forgetting irrelevancies) are crucial to keep the context useful indefinitely[[58]](https://arxiv.org/html/2504.07952v1#:~:text=in,Roediger%20%26%20Butler%2C%202011%3B%20Karpicke). There is also the risk of **knowledge dilution or error accumulation**: unlike gradient-based learning which slowly adjusts weights by averaging over many examples, in-context learning is instantaneous and can be **misled by one or two bad examples or malicious inputs**. This raises security concerns (e.g. prompt injection attacks could trick a model into “learning” a false fact within a session). Research on **robustness in in-context learning** is nascent – for instance, techniques to quickly “unlearn” a misleading prompt have been proposed (so-called fast explicit *unlearning* in context). Evaluating an LLM’s *long-term consistency* is also tricky: how do we know if an agent that has been running for days with a memory module is still factual and not gradually drifting due to compounded errors? These questions define an important frontier for 2025–2026.

In summary, the latest advances show encouraging steps toward LLMs that can **dynamically update their knowledge and behavior** using the context alone. Through synthetic memory frameworks, models accumulate experience over multiple queries; through implicit mechanism insights, we understand and enhance how models adapt internally; via retrieval and external knowledge, models stay up-to-date and domain-specialized without weight changes; and with lifelong learning benchmarks, we chart the path to agents that **continuously learn in an open-ended fashion**. Achieving all this reliably will bring us closer to truly adaptive, continually learning AI systems.

## Summary of Major Contributions (2025–Present)

| **Contribution (Year)** | **Key Idea** | **Significance** |
| --- | --- | --- |
| **Dynamic Cheatsheet (Suzgun et al., 2025)**[[4]](https://arxiv.org/html/2504.07952v1#:~:text=Engineering%20and%20Physics%20problems,driven%20learning%20characteristic%20of%20human)[[7]](https://arxiv.org/html/2504.07952v1#:~:text=execution%2C%20but%20lacks%20a%20memory,memory%20augmentation%20and%20efficient%20retrieval) | Persistent *external memory* at inference (store and retrieve solutions/insights between queries). | Enabled test-time *learning from experience* without weight updates; doubled accuracy on math and QA tasks by recalling prior successes. |
| **MemoryBank (AAAI 2024)**[[11]](https://arxiv.org/html/2504.15965#:~:text=limitation%20of%20LLM%E2%80%99s%20context%20window%2C,61)[[59]](https://arxiv.org/html/2504.15965#:~:text=memory.%20For%20example%2C%20MemoryBank%C2%A0,the%20external%20world%2C%20allowing%20the) | Long-term memory module for LLM agents (stores dialogue history & facts with update rules). | Allowed *continual learning* of user profiles and world knowledge across sessions; introduced memory decay/refresh mechanisms for stability. |
| **Implicit Gradient Descent in Transformers (Dherin et al., 2025)**[[21]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=Recent%20work%20by%20Dherin%20et,Intuitively%2C%20we%20can%20imagine%20that)[[24]](https://arxiv.org/abs/2507.16003#:~:text=seen%20during%20training,update%20of%20the%20MLP%20layer) | Theory showing transformer blocks implement *implicit low-rank weight updates* during forward pass. | Provided a mechanistic explanation for in-context learning; suggests LLMs literally **adjust internal weights** in response to prompts (without changing parameters), bridging prompting and learning. |
| **ICL as Associative Memory (Wu et al., 2025)**[[29]](https://icml.cc/virtual/2025/poster/44826#:~:text=%E2%80%9Clearn%E2%80%9D%20so%20quickly%20without%20updating,context%20learning%20is)[[30]](https://icml.cc/virtual/2025/poster/44826#:~:text=process%20can%20be%20understood%20as,depending%20on%20the%20question%20they%E2%80%99re) | Interpreted in-context learning as Hopfield associative memory retrieval conditioned on prompt examples. | Offered a new theoretical lens: prompt examples **reshape the model’s memory** to retrieve relevant patterns; explained why similarity and accuracy of examples matter for effective ICL. |
| **Attention Residual Stream (Burns et al., 2025)**[[60]](https://openreview.net/forum?id=lcTFm4LIRR#:~:text=computational%20neuroscience%20community%20to%20model,values%2C%20with%20results%20also%20indicating) | Novel transformer architecture allowing direct information flow between attention heads (inspired by neuroscience). | Achieved faster and stronger in-context learning in small models (8M–1B); demonstrates that architectural tweaks can improve *few-shot adaptation* at smaller scales. |
| **RARE – Retrieval-Augmented Reasoning Enhancement (Wang et al., 2025)**[[41]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=Enter%20RARE%3A%20A%20New%20Way,Forward)[[45]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=%2A%20Better%20Performance%3A%20A%20RARE,without%20retraining%20the%20whole%20AI) | Training paradigm that *splits knowledge and reasoning*: uses external knowledge base + trains LLM solely to reason with retrieved info. | Small (7–8B) models with RARE *outperformed GPT-4* on domain tasks; memory is external for easy updates, model excels at logical integration of facts (solving “hallucination” and poor reasoning). |
| **Dynamic Context Tuning (DCT, 2025)**[[48]](https://arxiv.org/html/2506.11092v1#:~:text=To%20address%20these%20issues%2C%20we,DCT%20introduces%20three%20key%20innovations)[[52]](https://arxiv.org/html/2506.11092v1#:~:text=We%20evaluate%20DCT%20on%20tasks,constrained%20settings) | Framework for multi-turn assistants: maintains a *context cache*, compresses history, and adapts a retriever (via LoRA) to new tools/domains on the fly. | Improved multi-turn dialogue consistency and tool-use adaptability (14% higher task success, 37% fewer hallucinations); shows value of combining short-term memory + adaptive retrieval for dialog agents. |
| **Lifelong Learning LLM Agents (Zheng et al., 2025)**[[19]](https://arxiv.org/abs/2501.07278#:~:text=This%20survey%20is%20the%20first,roadmap%20for%20researchers%20and%20practitioners) | Survey & roadmap categorizing modules (Perception, Memory, Action) for LLM-based agents to learn continually. Also introduced LifelongAgentBench evaluation. | Codified best practices and challenges for continual in-context learning; spurred development of benchmarks to measure an agent’s ability to *accumulate knowledge* without forgetting (stability-plasticity balance). |

**Table:** *Select advances in enabling LLMs to learn new information within their context window (2025+). These methods span persistent memory augmentation, theoretical insights into transformers’ implicit learning, retrieval and knowledge integration techniques, and frameworks for continual learning. Collectively, they push LLMs toward more* *adaptive, knowledge-updatable* *systems without altering core model weights.*

[[1]](https://arxiv.org/abs/2507.16003#:~:text=,reason%20why%20LLMs%20can%20learn) [[23]](https://arxiv.org/abs/2507.16003#:~:text=learn%20new%20patterns%20without%20any,Specifically%2C%20we%20show%20under%20mild) [[24]](https://arxiv.org/abs/2507.16003#:~:text=seen%20during%20training,update%20of%20the%20MLP%20layer) [2507.16003] Learning without training: The implicit dynamics of in-context learning

<https://arxiv.org/abs/2507.16003>

[[2]](https://icml.cc/virtual/2025/poster/44826#:~:text=Large%20language%20models%20like%20ChatGPT,depending%20on%20the%20question%20they%E2%80%99re) [[28]](https://icml.cc/virtual/2025/poster/44826#:~:text=Abstract%3A%20We%20provide%20an%20exactly,context%20examples%20shape) [[29]](https://icml.cc/virtual/2025/poster/44826#:~:text=%E2%80%9Clearn%E2%80%9D%20so%20quickly%20without%20updating,context%20learning%20is) [[30]](https://icml.cc/virtual/2025/poster/44826#:~:text=process%20can%20be%20understood%20as,depending%20on%20the%20question%20they%E2%80%99re) [[31]](https://icml.cc/virtual/2025/poster/44826#:~:text=process%20can%20be%20understood%20as,drawn%20from%20a%20familiar%20setting) [[32]](https://icml.cc/virtual/2025/poster/44826#:~:text=like%20how%20a%20person%20recalls,drawn%20from%20a%20familiar%20setting) ICML Poster In-Context Learning as Conditioned Associative Memory Retrieval

<https://icml.cc/virtual/2025/poster/44826>

[[3]](https://arxiv.org/html/2504.07952v1#:~:text=Despite%20their%20impressive%20performance%20on,truth%20labels%20or%20human) [[4]](https://arxiv.org/html/2504.07952v1#:~:text=Engineering%20and%20Physics%20problems,driven%20learning%20characteristic%20of%20human) [[5]](https://arxiv.org/html/2504.07952v1#:~:text=without%20retaining%20insights%20from%20previous,retaining%20algebraic%20insights%20across%20questions) [[6]](https://arxiv.org/html/2504.07952v1#:~:text=challenges%2C%20DC%20yields%20notable%20accuracy,responses%20and%20cutting%20routine%20errors) [[7]](https://arxiv.org/html/2504.07952v1#:~:text=execution%2C%20but%20lacks%20a%20memory,memory%20augmentation%20and%20efficient%20retrieval) [[8]](https://arxiv.org/html/2504.07952v1#:~:text=An%20even%20more%20striking%20and,manually%20fumbling%20with%20numeric%20manipulations) [[9]](https://arxiv.org/html/2504.07952v1#:~:text=We%20also%20note%20that%20DC,Roediger%20%26%20Butler%2C%202011%3B%20Karpicke) [[10]](https://arxiv.org/html/2504.07952v1#:~:text=thereby%20facilitating%20meta,data%2C%20results%2C%20and%20code%20at) [[20]](https://arxiv.org/html/2504.07952v1#:~:text=approach%20instead%20of%20manually%20fumbling,with%20numeric%20manipulations) [[58]](https://arxiv.org/html/2504.07952v1#:~:text=in,Roediger%20%26%20Butler%2C%202011%3B%20Karpicke) Dynamic Cheatsheet: Test-Time Learning with Adaptive Memory

<https://arxiv.org/html/2504.07952v1>

[[11]](https://arxiv.org/html/2504.15965#:~:text=limitation%20of%20LLM%E2%80%99s%20context%20window%2C,61) [[12]](https://arxiv.org/html/2504.15965#:~:text=task%20adaptability%20and%20contextual%20awareness,like) [[13]](https://arxiv.org/html/2504.15965#:~:text=match%20at%20L759%20MemoryBank%C2%A0,based%20retrieval) [[14]](https://arxiv.org/html/2504.15965#:~:text=limitation%20of%20LLM%E2%80%99s%20context%20window%2C,21%5D%2C%20etc) [[15]](https://arxiv.org/html/2504.15965#:~:text=I%20%20%20Personal%20Non,42) [[16]](https://arxiv.org/html/2504.15965#:~:text=Retrieval%20%20%20RET,124) [[17]](https://arxiv.org/html/2504.15965#:~:text=executing%20tools%2C%20thereby%20efficiently%20completing,addition%2C%20many%20commercial%20and%20open) [[59]](https://arxiv.org/html/2504.15965#:~:text=memory.%20For%20example%2C%20MemoryBank%C2%A0,the%20external%20world%2C%20allowing%20the) From Human Memory to AI Memory: A Survey on Memory Mechanisms in the Era of LLMs

<https://arxiv.org/html/2504.15965>

[[18]](https://arxiv.org/abs/2501.07278#:~:text=core%20components%20of%20these%20agents,roadmap%20for%20researchers%20and%20practitioners) [[19]](https://arxiv.org/abs/2501.07278#:~:text=This%20survey%20is%20the%20first,roadmap%20for%20researchers%20and%20practitioners) [2501.07278] Lifelong Learning of Large Language Model based Agents: A Roadmap

<https://arxiv.org/abs/2501.07278>

[[21]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=Recent%20work%20by%20Dherin%20et,Intuitively%2C%20we%20can%20imagine%20that) [[22]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=every%20prompt%20example%20attaches%20a,instructions%20that%20alter%20its%20computations) [[25]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=viewed%20as%20performing%20a%20form,disappear%20and%20the%20base%20model) [[26]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=prompt%20examples%20as%20if%20it,For%20example%2C%20even%20when%20prompt) [[27]](https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489#:~:text=The%20work%20of%20Dherin%20et,focus%20on%20how%20the%20self) Learning without Retraining: Implicit In-Context Learning Dynamics for Adaptive Cryptocurrency Systems | by Jung-Hua Liu | Jul, 2025 | Medium

<https://medium.com/@gwrx2005/learning-without-retraining-implicit-in-context-learning-dynamics-for-adaptive-cryptocurrency-661ffeeaa489>

[[33]](https://openreview.net/forum?id=lcTFm4LIRR#:~:text=computational%20neuroscience%20community%20to%20model,larger%20and%20more%20naturalistic%20scales) [[34]](https://openreview.net/forum?id=lcTFm4LIRR#:~:text=ICL,values%2C%20with%20results%20also%20indicating) [[35]](https://openreview.net/forum?id=lcTFm4LIRR#:~:text=LLMs,8%20million%20and%201%20billion) [[36]](https://openreview.net/forum?id=lcTFm4LIRR#:~:text=ICL,larger%20and%20more%20naturalistic%20scales) [[60]](https://openreview.net/forum?id=lcTFm4LIRR#:~:text=computational%20neuroscience%20community%20to%20model,values%2C%20with%20results%20also%20indicating) Associative memory inspires improvements for in-context learning using a novel attention residual stream architecture | OpenReview

<https://openreview.net/forum?id=lcTFm4LIRR>

[[37]](https://research.ibm.com/blog/larger-context-window#:~:text=Why%20larger%20LLM%20context%20windows,more%20coherent%20and%20relevant%20answers) Why larger LLM context windows are all the rage - IBM Research

<https://research.ibm.com/blog/larger-context-window>

[[38]](https://magazine.sebastianraschka.com/p/llm-research-papers-2025-list-one#:~:text=%2A%2023%20May%2C%20QwenLong,17667) LLM Research Papers: The 2025 List (January to June)

<https://magazine.sebastianraschka.com/p/llm-research-papers-2025-list-one>

[[39]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=Existing%20fixes%20have%20their%20own,on%20how%20to%20use%20it) [[40]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=A%20team%20of%20researchers%20from,it%20down%20in%20simple%20terms) [[41]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=Enter%20RARE%3A%20A%20New%20Way,Forward) [[42]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=outsources%20knowledge%20storage%20and%20focuses,level%20thinking%20over) [[43]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=,bogged%20down%20memorizing%20facts%2C%20it) [[44]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=1.%20High,world%20challenges) [[45]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=%2A%20Better%20Performance%3A%20A%20RARE,without%20retraining%20the%20whole%20AI) [[46]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=That%E2%80%99s%20a%20smaller%20AI%20outperforming,without%20retraining%20the%20whole%20AI) [[47]](https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c#:~:text=The%20RARE%20framework%20shows%20that,to%20make%20AI%20an%20expert) RARE: New Method to Replace RAG. Retrieval-Augmented Reasoning Modeling | by BOUKOUFFALLAH ABDALLAH | Medium

<https://medium.com/@la_boukouffallah/rare-new-method-to-replace-rag-5cfdd6a7e01c>

[[48]](https://arxiv.org/html/2506.11092v1#:~:text=To%20address%20these%20issues%2C%20we,DCT%20introduces%20three%20key%20innovations) [[49]](https://arxiv.org/html/2506.11092v1#:~:text=Multi,turn%20understanding) [[50]](https://arxiv.org/html/2506.11092v1#:~:text=Moreover%2C%20DCT%20generalizes%20to%20previously,wide%20range%20of%20dynamic%20environments) [[51]](https://arxiv.org/html/2506.11092v1#:~:text=On%20the%20tool%20adaptation%20front%2C,without%20retraining%20the%20base%20model) [[52]](https://arxiv.org/html/2506.11092v1#:~:text=We%20evaluate%20DCT%20on%20tasks,constrained%20settings) [[53]](https://arxiv.org/html/2506.11092v1#:~:text=Conversational%20RAG%20systems%20extend%20retrieval,not%20support%20active%20tool%20invocation) [[54]](https://arxiv.org/html/2506.11092v1#:~:text=Domain,8) Dynamic Context Tuning for Retrieval-Augmented Generation: Enhancing Multi-Turn Planning and Tool AdaptationThis research was conducted independently. No external funding was received.

<https://arxiv.org/html/2506.11092v1>

[[55]](https://www.nb-data.com/p/advancing-ai-lifelong-learning-roadmap#:~:text=1,knowledge%20when%20acquiring%20new%20information) [[56]](https://www.nb-data.com/p/advancing-ai-lifelong-learning-roadmap#:~:text=In%20contrast%20to%20traditional%20LLMs%2C,changing%20scenarios) [[57]](https://www.nb-data.com/p/advancing-ai-lifelong-learning-roadmap#:~:text=Large%20Language%20Models%20,their%20effectiveness%20in%20dynamic%20settings) Advancing AI: Lifelong Learning Roadmap for LLM-Based Agents

<https://www.nb-data.com/p/advancing-ai-lifelong-learning-roadmap>