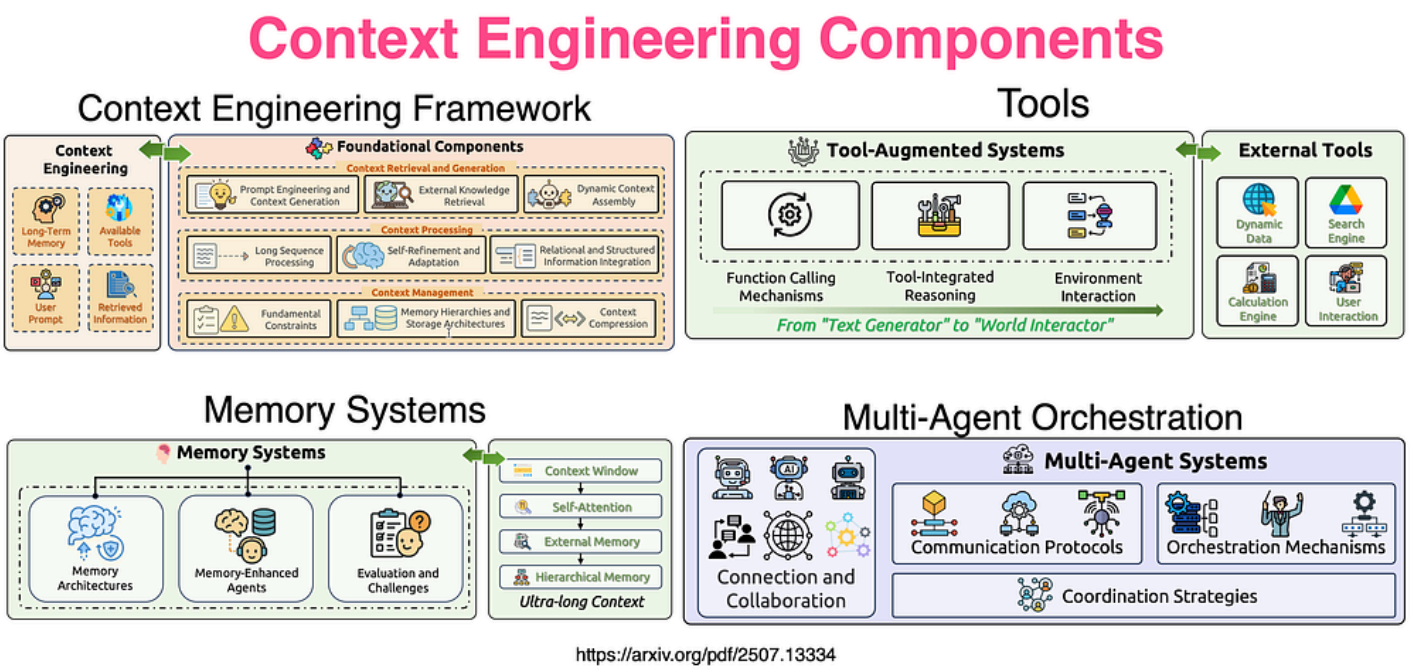
# Context Engineering

What are the components that make up context engineering & how can context engineering be scaled…

---

Author: Cobus Greyling | 6 min read | Aug 14, 2025
URL: https://cobusgreyling.medium.com/context-engineering-a34fd80ccc26

---




https://arxiv.org/pdf/2507.13334

The accuracy of Large Language Models (LLMs) is fundamentally determined by the availability of a contextual reference at inference.

A number of frameworks and methods have developed around scaling the establishment of accurate context at inference…RAG and others.

**Context Engineering** came to the fore as a new and format discipline that extends simple prompt design to include the process of systematically optimising the information supplied.

> Context Engineering combines techniques used to design, manage & optimise context.

I must say, this is a theme which has been emphasised by NVIDIA and named by them as the **data flywheel**…the continuous improvement of the AI Agent or inference by optimising inference based on input and output pairing.

A recent study define the taxonomy of this principle of Context Engineering…

# Foundational Components

A brief overview of the foundational components of Context Engineering…

### Context Retrieval & Generation

Prompt-based generation and external knowledge acquisition

### Context Processing

Addressing long sequence processing, self-refinement & structured information integration.

### Context Management

Covering memory hierarchies, compression & optimisation.

# System Implementations

### Retrieval-Augmented Generation (RAG)

Modular, agentic & graph-enhanced architectures

### Memory Systems

Enabling persistent interactions

### Tool-Integrated Reasoning

Function calling and environmental interaction

### Multi-Agent Systems

Coordinating communication and orchestration.

# The Evolution of Context

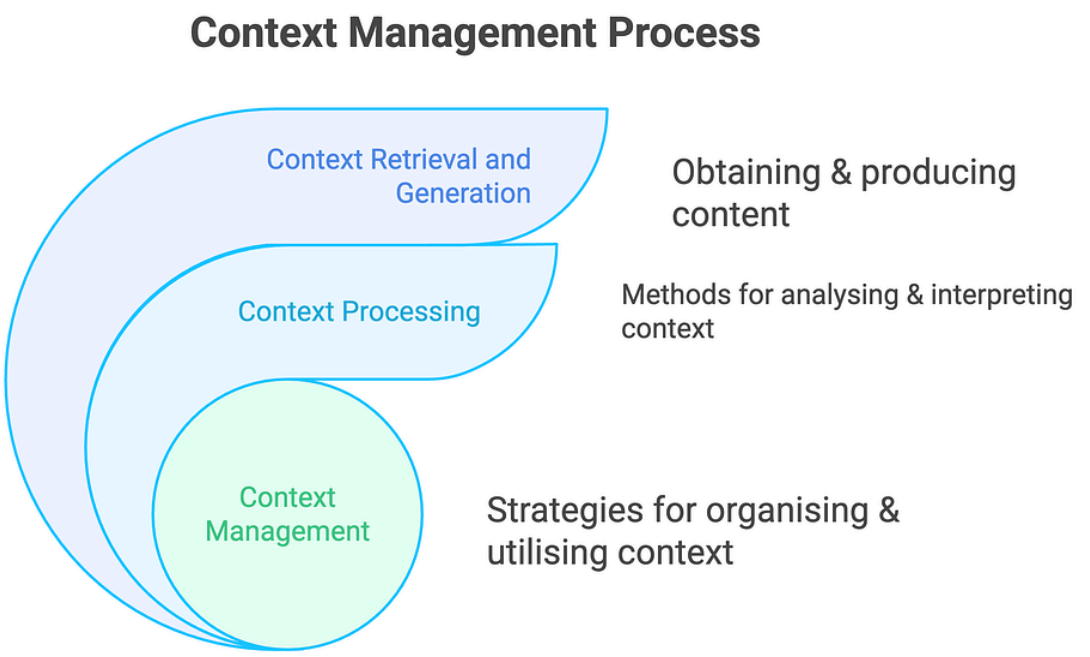
The evolution of context in **Large Language Models** follows a three-step framework known as **context engineering**, designed to enhance the robustness and efficiency of AI systems.

The first step, **Context Retrieval and Generation**, involves sourcing and creating pertinent information through methods like prompt engineering, external knowledge acquisition via retrieval-augmented generation (RAG), and dynamic context assembly to build a solid foundational layer for model inputs.

> Context enhances the robustness & efficiency of AI systems.

The second step, **Context Processing**, refines this raw context by techniques such as chunking, embedding, relevance scoring and compression to optimise for accuracy, reduce noise, and improve computational efficiency.

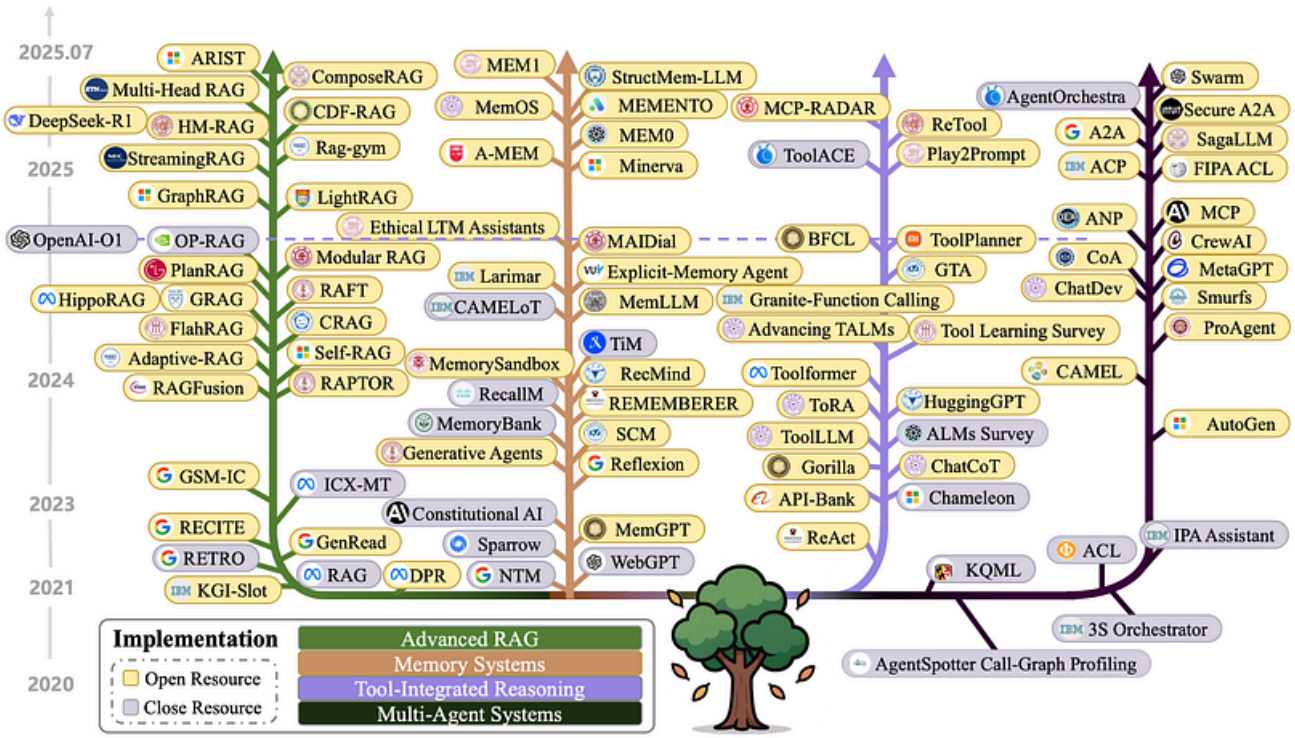
Finally, **Context Management** represents the advanced phase of ongoing oversight, where the system dynamically handles context through memory prioritisation, adaptation across interactions, and autonomous updates to ensure long-term coherence and relevance in complex, multi-turn scenarios.




The challenge is that many of the technologies have developed in isolation and the landscape is fragmented; one of the objectives and challenges of context engineering is to orchestrate or combine these technologies.

## Context Engineering Evolution Timeline

From the study, the timeline below shows the four branches of Context Engineering in the form of Advanced RAG, Memory Systems, Tool-Integration via AI Agents and Multi-Agent Systems and orchestration.




https://arxiv.org/pdf/2507.13334

## Context Scaling

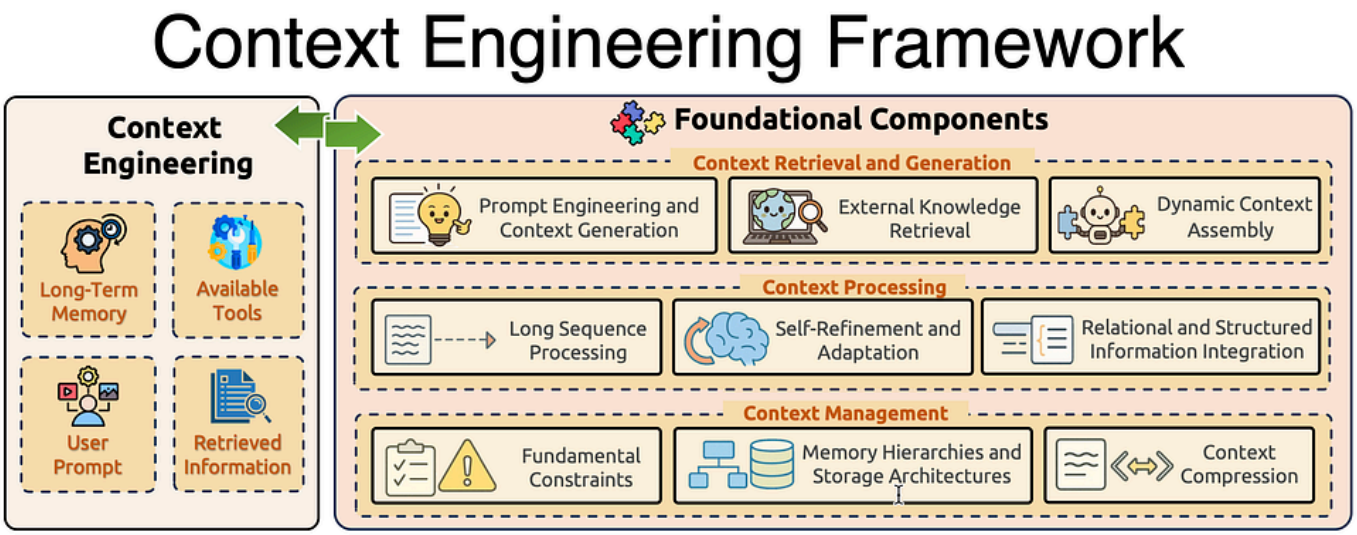
Creating context for individual, handcrafted cases is straightforward, but the real challenge lies in scaling it up — and developing mechanisms to do so effectively.

Context scaling involves two key aspects that determine how well systems can handle and process contextual information.

The first is **length scaling**, which tackles the technical hurdles of dealing with extremely long inputs.

This means expanding context windows from thousands to millions of tokens, while keeping the understanding coherent across long stories, documents, or conversations.

It relies on advanced attention methods, memory techniques, and model designs to manage this extended information without losing track.




https://arxiv.org/pdf/2507.13334

The second aspect is **multi-modal and structural scaling**, which goes beyond plain text to include richer, more varied types of context.

This covers things like temporal context *(time-based relationships and sequences)* spatial context *(location and geometry)*, participant states *(tracking multiple people or entities and their changes)*, intentional context (goals, motivations, and hidden objectives) and cultural context *(social and cultural nuances in communication)*.

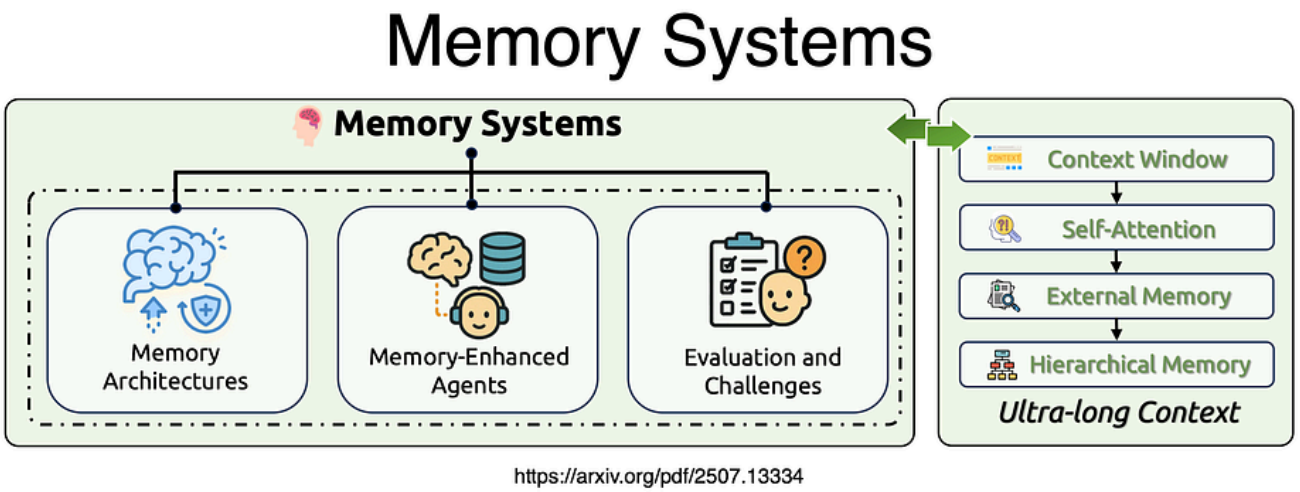
## Memory Systems

As mentioned, AI Agents overcome limitations in their built-in memory, like short context windows, by using advanced technologies such as Retrieval-Augmented Generation (*RAG*), which blends the agent's internal knowledge with external data sources without needing retraining.

> Memory enhancements let AI Agents handle complex tasks by integrating planning, tool use and multi-step reasoning through natural language.

This allows quick access to vast information through tools like vector databases for short-term contextual recall and long-term storage.

Non-parametric methods keep the core AI model unchanged while pulling in relevant details from outside.

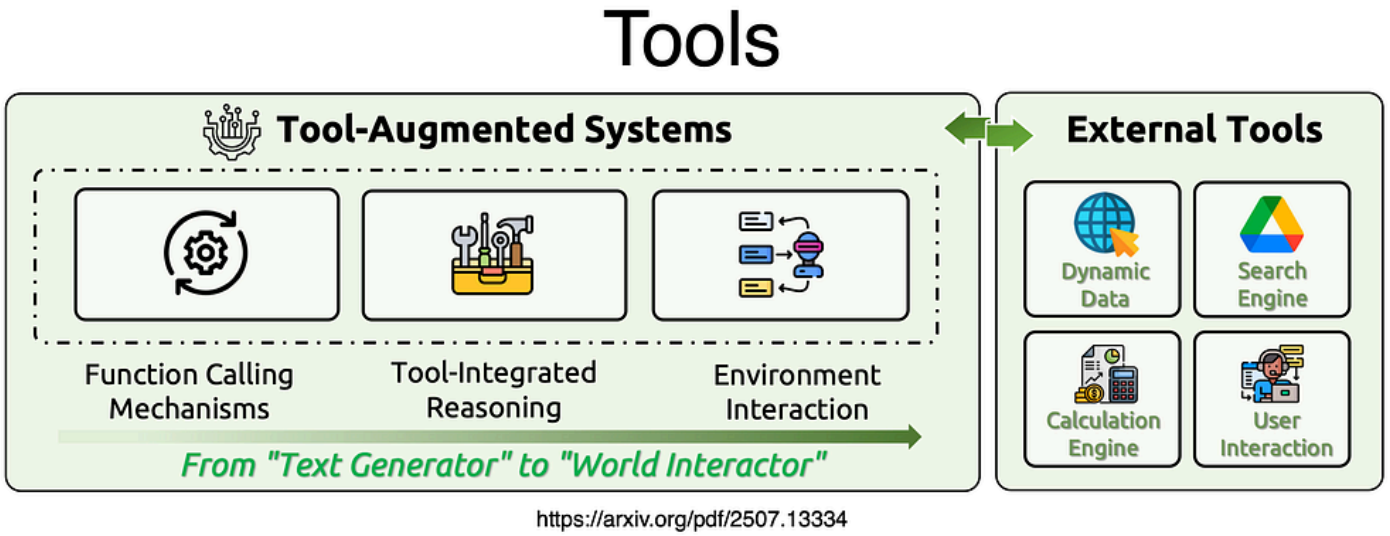



More sophisticated setups, with graph databases, organise memories for better retrieval, while commercial tools such as OpenAI's ChatGPT Memory and Apple's Personal Context enable personalised interactions.

Overall, memory enhancements let AI Agents handle complex tasks by integrating planning, tool use, and multi-step reasoning through natural language.

## Tools

Tools in the context of AI Agents refer to external functionalities or APIs that enhance an agent's capabilities beyond its core language model, enabling it to interact with the real world, retrieve information, or perform complex tasks autonomously.




Tools can range from web search engines and code interpreters to data analysers and image processors, allowing agents to fetch real-time data, execute computations, or manipulate media as needed.

By integrating tools, AI Agents become more versatile, transforming from passive responders into proactive problem-solvers that can chain actions — such as searching for facts, verifying them through multiple sources and generating outputs based on synthesised insights.

> Ultimately, tools empower AI Agents to bridge the gap between digital intelligence and practical utility.

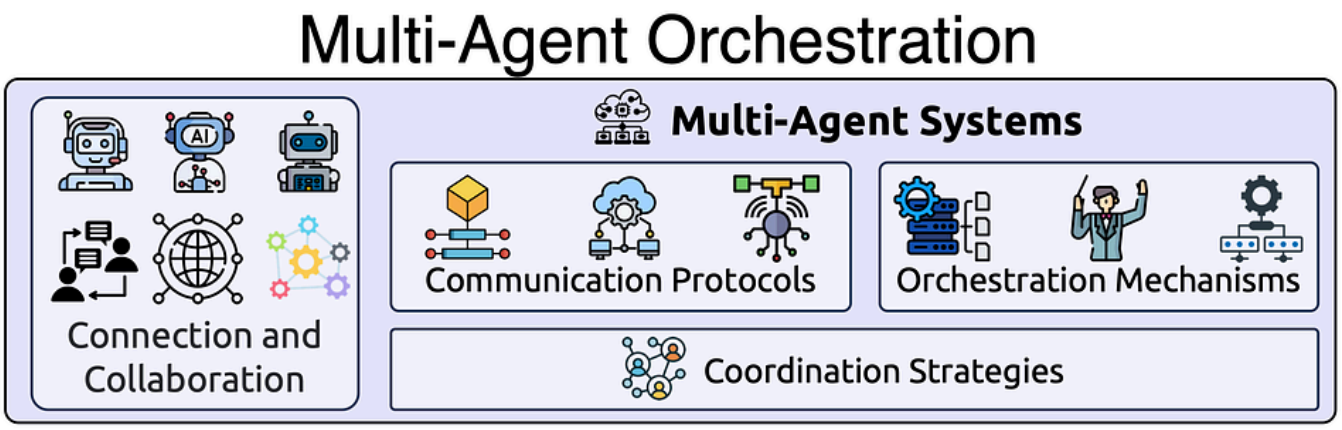
A tool-augmented approach mitigates limitations like knowledge cutoffs or hallucinations, for more accurate and efficient decision-making in applications like virtual assistants, automated research, or creative workflows.

Ultimately, tools empower AI Agents to bridge the gap between digital intelligence and practical utility, evolving them into sophisticated systems capable of handling multifaceted queries with precision and adaptability.

## Multi-Agent Orchestration

Multi-Agent Systems represent the pinnacle of collaborative intelligence, enabling multiple autonomous agents to coordinate and communicate for solving complex problems beyond individual agent capabilities.

> Multi-Agent Systems represent the pinnacle of collaborative intelligence.




https://arxiv.org/pdf/2507.13334

This implementation focuses on sophisticated communication protocols, orchestration mechanisms and coordination strategies that enable seamless collaboration across diverse agent architectures.

---

## Mentions by Author

- arxiv.org - A Survey of Context Engineering for Large Language Models - The performance of Large Language Models (LLMs) is fundamentally determined by the contextual information provided…
- Where AI Meets Language | Language Models, AI Agents, Agentic Applications, Development Frameworks & Data-Centric…
- github.com - GitHub - Meirtz/Awesome-Context-Engineering: 🔥 Comprehensive survey on Context Engineering: from… - 🔥 Comprehensive survey on Context Engineering: from prompt engineering to production-grade AI systems. hundreds of…