

AI Factory Stack



The **AI Factory Stack** illustrates the key components required to transform a general-purpose Large Language Model (LLM) into a reliable, production-ready AI system. Each layer adds a specific capability, moving from the core intelligence of the model to the infrastructure, governance, and operational mechanisms needed to deliver real-world AI applications.

1. LLM — The Machine

The Large Language Model (LLM) is the core intelligence of the AI Factory Stack. It is responsible for understanding language, generating responses, reasoning over information, and creating content. An LLM can answer questions, summarize documents, write content, and support conversations, but on its own, it is limited to the knowledge available to it and cannot reliably access private, current, or organizational information. It therefore serves as the foundational “machine” upon which the rest of the AI system is built.

2. RAG — The Raw Material Room

Retrieval-Augmented Generation (RAG) provides the LLM with access to external knowledge and organizational information. Instead of relying only on what the model learned during training, RAG retrieves relevant documents, policies, records, or other data sources at the time a question is asked. This retrieved information is then provided to the LLM to generate a more accurate and contextually relevant response. RAG effectively gives an AI system access to the knowledge it needs to perform specialized tasks.

3. Vector Database — The Storage Warehouse

The Vector Database stores information in a format that allows AI systems to search based on meaning rather than just exact keywords. Documents and other content are converted into numerical representations called embeddings, which capture their semantic meaning. When a user asks a question, the system searches the vector database for information that is conceptually related to the query. This makes it a critical component of RAG systems, enabling fast and intelligent retrieval of relevant knowledge.

4. AI Agent — The Floor Manager

The AI Agent acts as the decision-making and coordination layer of the system. While an LLM primarily generates responses, an agent can determine what steps are required to achieve a specific goal. It can plan tasks, decide which tools or information sources to use, execute actions, and evaluate the results before moving to the next step. In this sense, the AI agent acts like a floor manager, coordinating different components of the AI system to complete more complex, multi-step tasks.

5. MCP — The Power Outlet Standard

The Model Context Protocol (MCP) provides a standardized way for AI systems to connect with external tools, applications, databases, and services. Instead of creating a separate custom integration for every tool, MCP establishes a common interface through which AI agents and models can access capabilities such as APIs, databases, files, and enterprise systems. This makes AI applications more modular, interoperable, and easier to expand as new tools and services are added.

6. Guardrails — Safety Systems

Guardrails provide the controls and protections needed to ensure that AI systems operate safely and responsibly. They can help prevent harmful or inappropriate outputs, protect sensitive information, enforce organizational policies, and restrict the actions an AI agent is allowed to perform. Guardrails are particularly important when AI systems interact with real users, confidential data, or external systems. They ensure that increased AI capability and autonomy are balanced with security, governance, and human-defined boundaries.

7. Evals — Quality Check

Evaluations, or Evals, are used to measure how well an AI system performs against defined quality standards. They test factors such as accuracy, relevance, reliability, safety, consistency, and task completion. Rather than assuming that an AI application is working correctly, Evals provide a structured way to test its performance using representative scenarios and benchmarks. They are essential for identifying weaknesses, comparing changes, and ensuring that improvements to the AI system do not introduce new problems.

Overall, the AI Factory Stack shows that a successful AI product is much more than an LLM. The LLM provides the intelligence, RAG and vector databases provide knowledge, agents provide decision-making and action, MCP enables connectivity, guardrails provide safety, and evaluations ensure quality. Together, these layers form the foundation for building AI systems that are useful, reliable, scalable, and ready for real-world deployment.

Revision #1

Created 2026-08-24 11:56:25 UTC by Aisha M. Nur

Updated 2026-08-24 11:57:25 UTC by Aisha M. Nur