LUISS Guido Carli University

MSc in Data Science and Management

Course: Data Visualization

ENHANCING KNOWLEDGE ACCESSIBILITY FROM SHIFT DOCUMENTATION IN PHARMA MANUFACTURING: AN AI-BASED, HUMAN-CENTERED SOLUTION

Supervisor:

Prof. Blerina Sinaimeri

Co-supervisor:

Prof. Lorenza Morandini

Candidate:

Ludovica Autorino

Student ID: 783051

Academic Year 2024–2025

Acknowledgements

Alla mia famiglia:

Martina, Marco, Pietro, Camilla, Margherita, Josef, Ida e Camille.

Grazie per essere casa sempre, anche se distanti.

Grazie per essere sempre casa, anche quando siete lontani. Grazie per avermi sostenuta nelle mie scelte accademiche e professionali, anche quando diverse dalle vostre e talvolta difficili da comprendere. Grazie perché il vostro sguardo su di me non è mai mancato. Non sono più soltanto la piccola di casa, da sempre coccolata, ma una voce unica: ascoltata e valorizzata nella nostra grande, a volte caotica e rumorosa, famiglia.

Alla mia comunità:

Amici e Amiche di sempre o incontrati da poco. Con voi ho sperimentato la spontaneità e la bellezza di essere me stessa, ho imparato ad accogliere le mie vulnerabilità e ho scoperto che spendersi per l'altro è il vero Amore e il senso profondo della vita.

Ai miei colleghi e colleghe di Firenze:

Un piccolo capitolo della mia vita che però mi ha lasciato tanto.

Francesco, Alessandro, Elisabetta, Giuseppe, Marco, Alessandro, Matilde, Laura, e Irene.

Dal primissimo istante in questa città e in questa azienda mi avete fatta sentire accolta. Porto nel cuore ogni risata, chiacchiera e scambio che abbiamo condiviso. Un grazie speciale a Francesco e Alessandro per avermi seguita e sostenuta dall'inizio

alla fine di questo progetto: mi avete insegnato tantissimo, mi avete valorizzata e lasciata libera di esprimere le mie capacità e le mie difficoltà, anche nei momenti di sconforto in cui non mi sentivo all'altezza. Tutto ciò che ho imparato lo porterò con me ovunque.

Alla Prof.ssa Sinaimeri:

Per aver accolto con entusiasmo e curiosità la mia proposta di tesi fin dal primo momento.

Un grazie a tutti e tutte voi che ho incontrato nel mio percorso accademico, e a chi incontrerò nella strada che mi resta da percorrere...

Contents

Chapter 1

Introduction

Knowledge management is a crucial challenge in industrial settings, which is increasingly oriented towards efficiency and digitalization. If handled efficiently it becomes a special ingredient which really boosts the performance of processes in industrial settings. Including new technologies as Artificial Intelligence systems is one of the options the current state of the art proposes to tackle knowledge management problems in large organizations.

In industrial production floors, characterized by shift and team rotation during a working day, communication and efficient information communication between teams represent a critical node, frequently underestimated. The so-called "*shift handover*" are key documents in this process: they contain both structured data (e.g., batch number, quantity produced, scrapes) and unstructured notes pertaining to relevant events, anomalies, alarms or maintenance interventions.

In practice, these information remain frequently buried inside documents, whose accessibility is currently limited, as the users interested in this information can only retrieve it through their personal email inboxes, without a centralized access point that enables efficient consultation. This presents an obstacle to workers responsible of the different production functions, which are obliged to look for precious information manually, relying on memory or vague time references.

The objective of this thesis project, developed in collaboration with a pharmaceutical company, is the realization of an intelligent system which supports knowledge management in one of the production modules. The proposed solution leverages re-

cent advances in Artificial Intelligence and information retrieval to build an agentic architecture that enables users to efficiently search for, filter, and summarize the contents of historical shift reports.

Chapter 2 of this thesis underlines the contextual foundations underpinning this project. It begins by discussing the importance of knowledge management in pharmaceutical industrial contexts, highlighting how operational knowledge is often undocumented, scattered, or lost despite its value. The chapter then formalizes the problem statement and defines the objectives of the project. The lack of centralized searchability, the reliance on human memory, and the absence of a structured knowledge layer are framed as concrete barriers to cross-functional collaboration and process optimization. In response, the chapter outlines the overarching objective: to develop a modular and extensible system capable of retrieving, filtering, and summarizing shift handover information in response to user queries, thereby transforming raw documents into actionable insights. The chapter also defines the boundaries and assumptions of the project, such as the format and structure of the PDF reports, the confidentiality requirements of the data, and the practical considerations of deploying the system in a production setting.

Chapter 3 presents the state of the art technologies available and used to enable solutions as the one developed in this thesis. It again addresses the issues related to current knowledge management solutions. The limitations of current tools, are examined. It describes possible technologies used for effectively retrieve relevant information from a knowledge base. The chapter, subsequently, addresses ways to perform intelligent processing of information retrieved. Finally, it describes solutions to orchestrate different tools together, such as agentic artificial intelligence architecture.

A detailed description of the methodology followed to develop the final solution is provided by Chapter 4. It introduces LangGraph as the orchestration engine and details the agentic workflow: from user query intake (date range, keywords, production line), through hybrid retrieval over a Chroma-based vectorstore, a user selection interface, and finally the summarization module. Each component is discussed with attention to implementation choices, model selection, and data preparation strategies,

highlighting how the system maintains modularity, transparency, and scalability.

An extensive image of the output of the solution is given in Chapter 5, where a selection of use cases are presented and analyzed to validate the system's functionality. These use cases try to cover different realistic scenario of application of the proposed solution. The chapter illustrates how the system responds to user queries, showing sample outputs and discussing their relevance and coherence. The effectiveness of the system is evaluated qualitatively, based on its ability to retrieve meaningful content and generate informative summaries.

A more structured evaluation of the system is provided in Chapter 6. In this section first it is illustrated the choice made to evaluate such system. Namely, using a qualitative approach, comparing performance of this custom solution with a general-purpose commercial solution. Moreover a complete overview of the evaluation is discussed, considering limitations and further improvements of the system proposed.

Through this structure, the thesis aims to demonstrate how the integration of retrieval-augmented architectures and large language models can support the operationalization of knowledge in complex industrial contexts. By focusing on a real use case within a pharmaceutical production module, the work offers both a methodological contribution and a practical prototype. In addition to the system's technical description and testing, a critical reflection is also conducted, with particular attention to the implications of implementing such a solution within a highly regulated environment like pharmaceutical manufacturing. This includes considerations on data handling constraints, traceability needs, and system robustness. The discussion highlights the importance of developing tailored, human-centered solutions, designed around the actual workflows and information needs of end users. At the same time, it reflects on what such customization entails in terms of design effort, integration complexity, and required resources, especially when compared to more general-purpose or off-the-shelf systems. The following chapters guide the reader through the rationale, design, implementation, evaluation, and implications of the proposed solution, ultimately showing its potential to improve knowledge accessibility and decision-making on the production floor.

Chapter 2

Context Description and Problem Statement

This chapter introduces the problem that motivated the development of this thesis project: knowledge management in an Italian pharmaceutical industrial context. It starts with a contextual overview of knowledge management practices in industries, highlighting how they are addressed in current scenarios. Subsequently, it delves into a detailed examination of the specific problems approached by this work, particularly the complexities associated with managing and utilizing unstructured operational data which result from handling all the knowledge passed between one shift and another. Understanding the problems that this current management presents brings to the definition of a clear objective that this thesis project aims to tackle, situating the proposed solution within the problem stated and existing cutting edge technologies present at the stat of the art. It follows an in-depth examination of the documents taken into consideration and the related challenge they present within this scenario.

2.1 Description of the context

2.1.1 Knowledge Management (KM) in Manufacturing Organization

According to the *Cambridge Dictionary* (Cambridge University Press n.d.), *knowledge* is defined as:

> **knowledge**
>
> *noun*
>
> UK /'nol.ij/ US /'na:.lij/
>
> *understanding of or information about a subject that you get by experience or study, either known by one person or by people generally.* (Cambridge University Press n.d.)

What can be considered *knowledge* in a manufacturing context is hard to define. It may extend beyond technical data or operational instructions. Indeed, it's coverage comprehends a wide range of strategic, organizational, and cultural elements.

Shaw and Edwards (2006) synthesize foundational contributions from scholars such as Hayes and Wheelwright (1984) and Hill (1987) to propose a structured understanding of what manufacturing knowledge entails. The first authors describe the evolution of manufacturing strategy: from being internally neutral to becoming externally supportive. Saying this, they highlight that manufacturing strategy, overtime contributes systematically to the company's overall strategy, reaching an alignment with the latter. Supporting this view, Hill distinguishes two key components of manufacturing strategy: structure and infrastructure. The structure refers to the tangible, technical components of manufacturing, such as production processes and the technologies used. In contrast, the infrastructure encompasses longer-term, cross-functional elements like human resource policies, quality systems, organizational culture, and information technology. In this sense, infrastructure plays a crucial role in supporting and enabling the structure to function properly and evolve over time.

From this perspective a categorization of knowledge in a manufacturing context may comprehend: operational knowledge, which is more related to the daily execution of production tasks, manufacturing strategy knowledge, which is, on the other hand, related to processes, technologies, systems, policies and procedures, and corporate strategy knowledge, which, in conclusion, is what sets the direction of manufacturing in alignment with business goals (Shaw and Edwards 2006).

It emerges that manufacturing knowledge is not only the *know how* in pure production processes, but also the overall organization, improvement and alignment of production with business objective. Such knowledge is distributed and often difficult to formalize its management particularly challenging in complex industrial contexts.

The process of developing, storing, retrieving, and sharing knowledge and expertise inside an organization for the purpose of improving its business performance is known as knowledge management, or KM (Gupta, Iyer, and Aronson 2000). Businesses are coming to understand that knowledge, in any form, is an essential source of value and has to be handled carefully: to become innovative as well as to remain competitive. To succeed, KM necessitates a significant change in corporate culture and dedication from all firm levels (Gupta, Iyer, and Aronson 2000).

It may be identified a linkage between knowledge management and organizational performance, information technology, competitiveness, the transfer of best practice, inter-organizational networking and organizational learning (Shaw and Edwards 2006).

KM emerged with not only the need to be cost-efficient and managerially effective in problem solving, decision making, innovation and all other elements needed to maintain and develop a competitive edge, but also more specifically, to capture, catalog, preserve, disseminate the expertise and knowledge that are part of organizational memory that typically resides within the organization in an unstructured way (Gupta, Iyer, and Aronson 2000).

Among the potential benefits of an effective KM is the support the development of skills and competences for Industry 4.0 (Ribeiro et al. 2022). Hence, a structured KM can facilitate knowledge sharing between experts and novices operators during training or organizational learning (Ribeiro et al. 2022). At the same time, the

technological advancements brought by Industry 4.0—particularly the increased volume and ease of data processing enabled by advanced industrial machinery—serve as fundamental enablers of effective Knowledge Management. There is a mutual reinforcement between the two. However, in order to harness these opportunities without contributing solely to increased system complexity, coordinated efforts across all levels and functions of the industrial organization are essential (Ribeiro et al. 2022).

Manufacturing Knowledge Management Trends

In modern manufacturing environments, Knowledge Management (KM) plays a central role in ensuring operational continuity, quality, and innovation. Organizations today handle a wide variety of knowledge types, typically divided into explicit and tacit categories (Symestic GmbH n.d.). Explicit knowledge includes documented manufacturing instructions, standard operating procedures (SOPs), technical specifications, CAD models, and quality standards. Tacit knowledge, on the other hand, resides in the practical experience of operators and engineers: it comprehends troubleshooting expertise, intuitive process optimizations, and deep contextual understanding of production systems (Symestic GmbH n.d.).

To manage this complex knowledge landscape, manufacturing firms adopt structured KM frameworks composed of several core components: acquisition, organization, provision, transfer, and application. This includes processes such as capturing experiential knowledge from the shop floor, digitizing analog records, organizing documents by process or product area, and delivering relevant content directly to the workplace through digital platforms (Symestic GmbH n.d.).

Technologies such as Document Management Systems (DMS), Enterprise Content Management (ECM), and Manufacturing Execution System (MES) integration support these activities, along with enablers like semantic search, augmented reality (AR), and AI-powered knowledge discovery (Symestic GmbH n.d.). These tools and methods aim to foster knowledge sharing, reduce redundancy, and promote continuous improvement across manufacturing lines.

According to Gupta, Iyer, and Aronson (2000) a cultural change, management practices, and commitment by all levels of the organization have to be put in place

in order to make these technologies truly effective in a manufacturing organizational structure.

Despite the availability of increasingly sophisticated technologies, managing knowledge in manufacturing remains a complex organizational challenge. Information continues to grow at an exponential rate, and appears in multiple formats: from casual emails and instant messages to structured reports and digital dashboards. This fragmentation makes it difficult to transform knowledge into insights. In response, some companies have appointed a Chief Knowledge Officer (CKO) to oversee knowledge governance and foster a culture of sharing (Gupta, Iyer, and Aronson 2000). Yet, internal silos, limited cross-functional communication, and resistance to behavioral change often hinder the effectiveness of such initiatives. To truly leverage knowledge as a strategic asset, organizations must go beyond storing information: they must understand who holds the knowledge, how it circulates within and across teams, and which knowledge needs to be shared, with whom, and why. Creating effective linkages between structured and unstructured knowledge, that is anchored to specific problems, processes, or decision-making contexts, is essential for a mature and impactful KM strategy.

Moreover, while technological tools and frameworks offer a robust foundation for enterprise-wide KM, they often fall short when it comes to managing unstructured, context-rich operational knowledge. The following sections will show a real world example of a knowledge management problem that this thesis is entitled to address: industrial reports drafted during shift handover.

2.2 A Focus on Industrial Shift Handover Reports

2.2.1 Descritpion of Shift Handover in a Pharmaceutical Manufacturing Company

This thesis originates from the need to address a common challenge in industrial settings: knowledge management. The pharmaceutical manufacturing sector has been chosen as the context of application because it represents an environment where this challenge becomes particularly critical. Pharmaceutical production is characterized

by strict regulatory frameworks and the necessity to ensure the highest quality standards, since even minor deviations can compromise patient safety. This entails that every aspect of production must be carefully documented, monitored, and traceable. As a consequence, effective knowledge management is not only desirable but essential, both to guarantee compliance with regulatory requirements and to support operational efficiency. The thesis therefore develops and tests its approach within an Italian pharmaceutical manufacturing company that produces life-saving medicines on a large scale.

Production in industrial organization is mostly organized in shifts. These shifts are taken over by different teams of operators lead by a production supervisor. At the end of each shift an important knowledge transfer has to occur, which is all the information regarding what happened on the shop floor during the shift: frequent alarms, down times, particular problems with a particular machine, whatever maintenance has been made. This knowledge is commonly referred to as *shift handover* and it allows to have a smooth transition between one shift and another.

A great innovation has been implemented by digitalizing this information in semi-structured report. There has been a transition from having this knowledge transmitted orally between supervisors, which made it impossible to track and preserve information systematically, to the collection of written digital records, which was a first step in the management of this knowledge preventing it to be lost and making it available to a greater audience.

In a manufacturing environment, indeed, a vast amount of information is collected on a daily basis. While most of it consists of structured production data, there is also a considerable portion of knowledge which remains useless if not properly leveraged and utilized. This occurs despite its potential value in addressing common problems and, therefore, continuous improvement in efficient production. This is the case with the amount of documents called *shift handovers* produced at the conclusion of each manufacturing shift of a production line in a pharmaceutical manufacturing site.

It is important to emphasize this project's area of application since it serves as the foundation for the infrastructure it was built on. In order to convey crucial information to the incoming shift, engineers and technicians responsible for the production

processes of a given line currently rely on a written report sent by email, which summarizes the key issues and performance metrics from the shift of the previous day. These reports are known as *shift handovers* and include a wealth of information that, after reading, is currently useless and remains an untapped source of valuable and insightful knowledge.

This thesis project is the outcome of a deep analysis of the manufacturing context, the problems posed by the configuration and accessibility of these documents, potential exploitation of the information contained, and the related needs.

2.3 Objective

The objective of this thesis project is to provide a digital solution for the context mentioned above. The goal is therefore to find a way to make use of the information contained in these documents not only efficiently, but also in a natural, user-friendly, and seamless manner. The solution is designed to ensure that knowledge becomes accessible and usable interactively, allowing relevant features to be condensed through intelligent information processing. All of this must be achieved while maintaining a high level of reliability and precision in data handling, given the stringent regulatory requirements of the pharmaceutical manufacturing environment in which the solution will be deployed.

2.4 The user persona(s)

To better define the requirements of the solution, user persona(s) were created to represent the typical individual engaging with shift handover reports in their existing access configuration, common practices and typical usage pattern.

2.4.1 Who fills in the shift handover

The shift handovers are filled in at the end of every production shift by individuals responsible for the production line. These operators carry out a set of routine activities, often referring to machines using conventional or informal names that may be familiar within a specific team but unclear to other audiences.

These documents are multi-lingual, which reflects the informality commonly used in industrial knowledge management practices: there is both Italian, language spoken at the production site, and some English words, due to common industry conventions. The same applies to acronyms, which are sometimes unofficial, functioning more as pseudo-acronyms without a standardized reference. The language and the structure used are different from production line to production line, even if in the same organization. Moreover, as will be described later, these documents consist of two main parts: one composed of pre-filled performance indicators, and the other of free-text sections where the most relevant events of the shift are described—such as recurring alarms, encountered issues, machine down times, and the corresponding interventions.

2.4.2 The daily reader

The second user group that interacts with shift handovers includes all associate operational figures within the production team, such as engineers responsible for operations, mechanical processes, performance, quality, and other related areas. These individuals are highly interested in the information contained in shift handover reports, as they are accountable for the outcomes of production activities but are not physically present on the shop floor for all the production time. This user persona receives three shift handovers per day via automated email, corresponding to the three daily shifts. These documents remain stored in each user's personal email inbox, which constitutes the sole point of access. This type of user typically focuses on information relevant to their specific domain of responsibility and from the recent past. However, they do not usually consult these documents as tools for emergency response—rather, only when the issues reported reach a certain degree of significance or persist over time, and these documents may become useful in showing the evolution of those problems. These engineers operate under strict quality and precision standards typical of the pharmaceutical industry, where every piece of operational information can contribute to continuous improvement and more efficient problem solving.

2.4.3 The exceptional reader

The exceptional reader, which may coincide with the daily reader-particularly when the latter is required to investigate the evolution of a specific issue in more detail-is a user involved in activities that go beyond routine operational monitoring. Even if not the primary audience, this type of user may still need to consult shift handover documents for various purposes. This user is generally involved in the analysis of issues that have evolved over time and escalated into more complex systemic problems, maybe regarding the whole production line or module. Consequently, they demonstrate a stronger need for information that is consolidated, processed, and presented in a manner that facilitates the extraction of actionable insights. Their perspective is generally oriented toward broader business or process-level implications, rather than toward isolated mechanical or technical anomalies.

Below the image 2.1 exemplifies the interaction that users have with these documents.

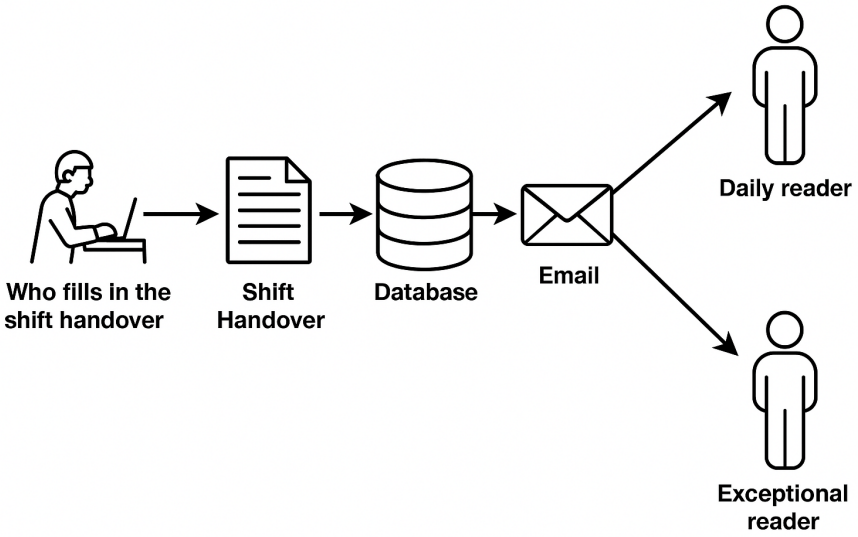

Figure 2.1: User diagram: how they interact with the shift handover today. Generated with DALL·E (OpenAI 2025).

2.5 Shift Handover Description

2.5.1 Document Structure

It is crucial to describe in detail how *shift handovers* are constructed to understand the considerations at the core of the decision making process throughout the development of the solution. Below in figure 2.2 is shown a template of the structure of these documents.

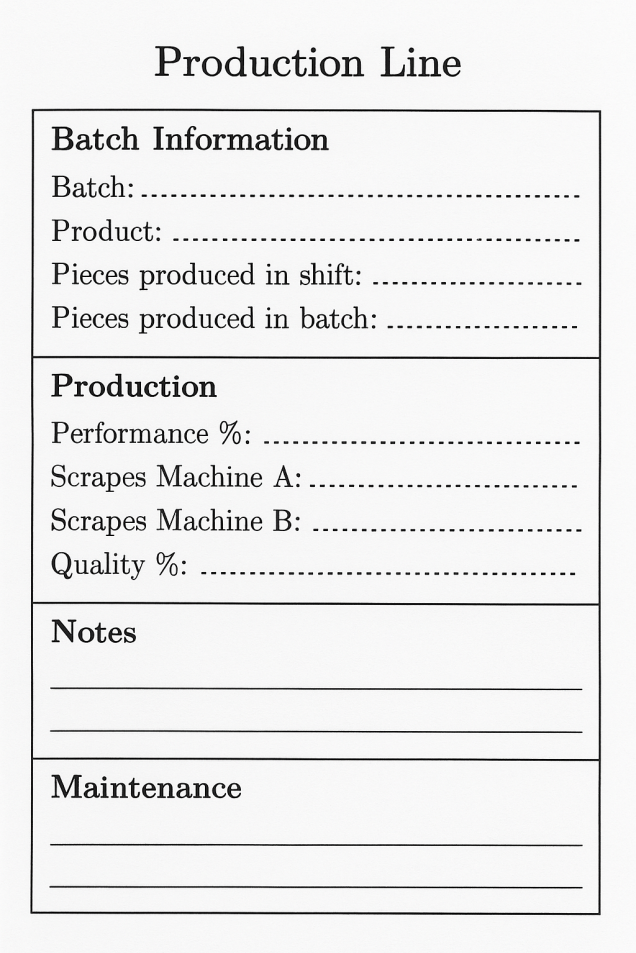

Figure 2.2: Shift Handover template. Generated with DALL·E (OpenAI 2025).

First of all, the documents are partially completed using a pre-filled template, which standardizes certain fields while leaving others to be manually completed by operators. The standardized fields are those regarding batch identification number, total produced in a batch or in the full shift, in cases of more than one batch per shift, percentage of scraps of the most important machines: in short, all common-

used performance indicators of the shift.

On the other hand, there is a section of the shift handover that has to be filled with text. This is the field intended to capture any noteworthy events occurred during the shift that require special attention. These may include description and explanation of production's downtimes due to specific alarms at a given station of a given machine, or it may include information regarding some mechanical interventions. Given that the first section is filled with quantitative data that can be eventually found elsewhere being logged automatically in other systems, these last information are the ones which remain lost and cannot be leveraged in the long term to analyze production trends and extract valuable operational insights.

Moreover, as briefly mentioned earlier, these documents are characterized by high variability and a risk of errors. This variability stems from the fact that different individuals, working across different shifts, are responsible for drafting them - often leading to typos and inconsistencies. In addition, each production line may adopt its own set of conventions, further contributing to the heterogeneity of the documentation.

On the other hand, the information contained is characterized by a certain degree of redundancy, making it difficult to extract useful insights. Indeed, the machines are not more than 20 per production modules, many of which share similar alarm names. Additionally, the textual content is typically brief and not very descriptive, resulting in essential and somewhat aseptic narratives that lack a comprehensive explanation of the issues encountered during the shift. Furthermore, the text is multilingual: primarily redacted in Italian, combined with conventional or technical names in English-not to mention the amounts of acronyms which may also vary across different documents.

This raises the first challenge: the attributes presented by the data to be processed. Namely the difficulty resides in developing a digital solution capable of capturing and processing this complex interplay of qualitative and quantitative data, characterized, on one hand, by variability and inconsistencies, and on the other, by redundancy and lack of descriptive content.

2.5.2 The role of Metadata

Metadata represent a key element of this documents. They are made of a combination of *Date*, *Shift* and *Production Line*. Metadata are data that describe other data. In this scenario, they represent essential information that enable the user to identify the events to which these documents refer and, therefore, must be preserved and transferred accordingly. Metadata are an instrument, from an archival point of view, to interact with these document. It is a knowledge that cannot be lost.

This poses the second challenge: to find a way to process these documents keeping a channel that allows these metadata not to be lost, and kept as source of reliability and accountability of the information processed in these documents.

2.5.3 Time coordinates

Together with metadata, time also plays a fundamental role in this context. The continuous production process generates a steady flow of information, with three different shift handovers produced every day for each line. Over weeks and months, this frequency leads to a rapidly growing archive of documents. This represents the third challenge to address: the volume of handovers makes scalability an issue for manual consultation, while at the same time the availability of such a rich temporal sequence of reports enables the recognition of recurring patterns and the evolution of specific production issues. This temporal information must therefore be leveraged rather than overlooked.

2.5.4 How to access and process this documents

The accessibility of knowledge contained in these documents for their intended audience is currently limited. In practice, access relies on the personal e-mail inboxes of the recipients, making it difficult to retrieve specific information without remembering the exact date or context. Moreover, keyword-based search is often ineffective, unless looking for very specific or unique terms, due to the high degree of redundancy in machine names and alarm descriptions, as will be discussed later.

Although documents are archived in a local database on premises accessible

through a virtual machine, managing and processing the data they contain can also be done via a cloud infrastructure, which supports real-time processing and improves the scalability and availability of the information.

Chapter 3

State of the Art

This chapter outlines the technological landscape relevant to the design and implementation of an advanced knowledge management (KM) solution for unstructured industrial data. Starting from the broader context of KM in operational environments, it then explores the core enabling technologies: semantic and keyword-based retrieval, Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), and agentic AI frameworks. These tools collectively provide the foundation for developing a customized architecture capable of addressing the unique challenges highlighted in Chapter 2.

3.1 Knowledge Management in Industrial Operations

As introduced in Chapter 2, industrial KM refers to the strategies, processes, and tools aimed at capturing, sharing, and leveraging knowledge across operational settings. In production contexts such as pharmaceutical manufacturing, this knowledge is not limited to procedures and formal documentation, but includes informal, tacit insights generated daily by operators, technicians, and supervisors.

A critical obstacle to effective KM in this domain is the dispersion of information. Industrial personnel frequently resort to ad-hoc solutions: spreadsheets, personal notebooks, email threads, and informal conversations to find or share critical information. Over-reliance on individuals' tacit knowledge is common. Shift handover reports, the object of this work, which often contain vital notes on machine anoma-

lies, interventions, and production status, are shared via email in unstructured or semi-structured formats. These reports vary in style and completeness, complicating knowledge extraction and reuse.

According to the International Data Corporation (IDC), employees spend an average of 2.5 hours a day looking for information, however they acknowledge that this is an estimate and might vary significantly depending on the type of job (IDC 2018). Existing digital solutions such as Manufacturing Execution Systems (MES), Document Management Systems (DMS), and collaborative platforms offer partial support. However, they typically lack the ability to process heterogeneous texts, perform semantic searches, or summarize domain-specific knowledge. In general there is a lack of a solution that integrates all these features in one single entry point.

3.2 Search Technologies in Knowledge Management

A backbone of any KM solution is the ability to efficiently search and retrieve relevant information. To solve daily issues, investigations in past incident logs, reports or instructions are frequent activities. To tackle this, traditional keyword-based search techniques and newer semantic search methods come into place.

3.2.1 Keyword-Based Search

Word-matching algorithms are the traditional approaches to information retrieval.In particular, among the most historically used models is the bag-of-words (BoW), a representation approach on which weighting schemes like Term Frequency-Inverse Document Frequency (TF-IDF) have been built to evaluate and rank document relevance. TF-IDF is a refinement of BoW that weighs words by their importance. The importance is measured by counting the word frequency in a document (TF), but dividing it by the frequency of the same words in common with other documents (IDF), helping to highlight more informative words (Manning, Raghavan, and Schütze 2008). The most used ranking solution for information retrieval, developed in the 90s, is the probabilistic algorithm BM25, which is an improved version of

TF-IDF, since it considers word frequency, document length, and saturation (diminishing returns of repeating a term), providing better retrieval performance, especially for longer documents or queries (Robertson and Zaragoza 2009).

BM25 is efficient for capturing all those documents that contain the exact query word, which is useful in industrial reports which are rich in code names, machines, dates, product or batch IDs. On the other hand, this kind of keyword-based search is not at all effective in capturing the semantic meaning of words and context: therefore, it is not ideal in cases where relevant information uses different wording. In a knowledge base of manufacturing reports, a user's query might not use the exact same terms as the document that contains the answer. As a result, purely keyword-based KM systems often have low recall for complex queries.

3.2.2 Semantic Search with Dense Embeddings

To fill this gap, a more recently used method is to employ a semantic search using dense vector embeddings (*Vector Stores | LangChain* 2024). In this way there is a shift from a keyword match to a semantic and conceptual match. It works by encoding textual data into high dimensional numeric vectors such that the distance between the vectors corresponds to the distance between the semantics of such words. This generates a vector space, which can be queried and it allows to retrieve documents conceptually related to the query even if they do not share literal keywords.

Recent progress in deep learning, especially transformer-based language models, has made it possible to create sentence and document embeddings that capture meaning in a rich and useful way. A key example is Sentence-BERT (Reimers and Gurevych 2019), which introduced a samese network architecture that fine-tunes BERT to generate sentence embeddings. This makes it much faster to compare sentence meanings than using raw BERT outputs (Reimers and Gurevych 2019). Models like SBERT and the Universal Sentence Encoder produce vector representations that can be stored in a vector database and quickly searched using nearest-neighbor methods.

Figure 3.1 represents how similarity search works: the vector store contains all the vectors of the textual data embedded, the query as well is embedded, using similarity

search between the query embedded and the vector store, relevant information is retrieved.

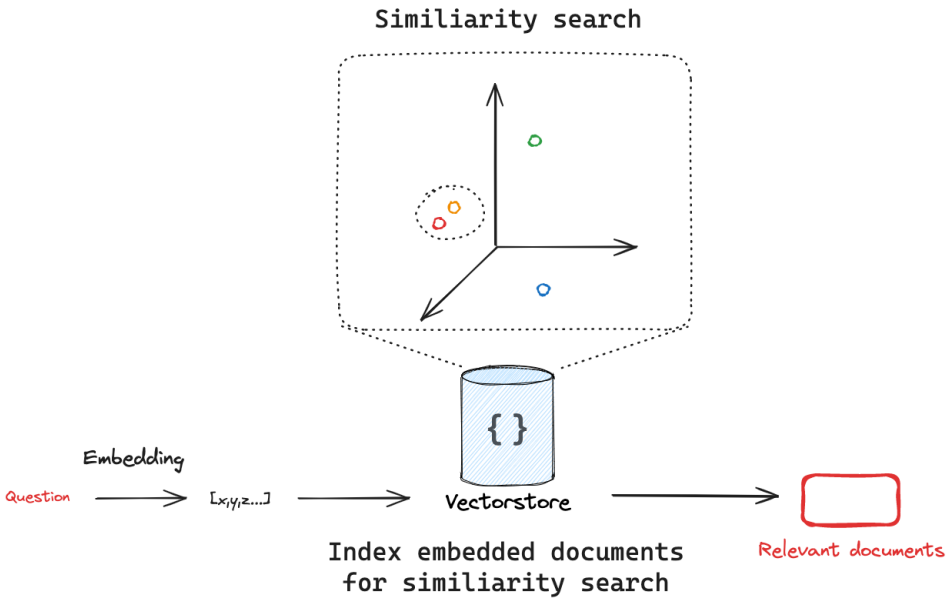

Figure 3.1: Graphical representation of how a vector store similarity search works (*Vector Stores | LangChain* 2024).

In practical applications, companies can rely on APIs that provide access to pre-trained embedding models, eliminating the need to develop and train models in-house. A notable example is OpenAI's `text-embedding-ada-002`, which generates 1536-dimensional vectors and has demonstrated strong performance on benchmarks for text search and semantic similarity (OpenAI 2022a).

This property enables semantic search systems to retrieve relevant documents even when they use different terminology than the user's query, which may result as a particularly valuable feature in fields like pharmaceuticals, where synonyms and domain-specific language are common. However, this comes at the potential cost of reduced precision: since retrieval is based on meaning rather than exact terms, some results may be conceptually related but not directly aligned with the user's intent.

3.2.3 Hybrid Retrieval Approaches

In high-stakes contexts, the trade-off between the two approaches described respectively in section 3.2.1 and 3.2.2 may require a solution that combines of both. This modern information retrieval architecture balances precision and contextual flexi-

bility, allowing to leverage both sparse and dense retrieval. You can combine the results from both keyword-based and semantic searches into a single ranked list to take advantage of both approaches, or use one method to filter candidates for the other, for example, retrieving a broad set of documents with keyword search and re-ranking them using semantic similarity. Alternatively, one can merge the relevance scores from BM25 and embedding-based similarity using a weighted combination, allowing fine-grained control over the influence of each retrieval method. In industrial KM, such a hybrid approach is ideal: if a user queries a batch number or uses a technical term, the keyword part guarantees those are present in results; meanwhile semantically similar content, perhaps using a different acronym or phrase, is not overlooked.

3.3 Large Language Models (LLMs) for Summarization

A great revolution of the last years in text summarization and understanding context and generate human-like language are Large Language Models (LLMs). Among those, remarkable capabilities in condensing documents, extracting key points and interpreting data there are the Generative pre-trained transformers (GPT). These are based on the transformer architecture, a deep neural network designed for natural language processing tasks and due to their remarkable performance on NLP (Natural Language Processing) tasks have gained traction among the researchers and the industrial communities, Establishing them as some of the most commonly adopted and powerful models in NLP and adjacent domains (Yenduri et al. 2023).

3.3.1 Capabilities and Industrial Use Cases

Brown and al. (2020) write that with OpenAI's GPT-3, The focus in NLP has transitioned from developing representations and models tailored to specific tasks to employing pre-training and architectures that are general-purpose and adaptable across tasks. In 2023, OpenAI has delivered GPT-4 which is an evolved GPT solution, considered better in reliability, creativity and ability to handle more complex instructions (OpenAI 2023). A further step by OpenAI, was the launch in 2024 of

GPT-4o, which takes the best of the previous models adding multi-modality features (OpenAI 2024). Thanks to its enhanced language comprehension, it can integrate information from various sources and apply reasoning to generate a clear and compact summary. A major critical point of GPT-4 and GPT-4o and of earlier models is its risk of *hallucination* of facts or making errors (OpenAI 2024). Therefore, great care should be taken when using language model outputs, particularly in high-stakes contexts, with the exact protocol matching the needs of a specific use-case. For critical domains like pharma, ensuring the LLM only uses provided source data is essential. This is where prompt design and retrieval augmentation (next section) help constrain the model.

3.3.2 Prompt Engineering

Prompt engineering is a technique to influence the effectiveness of LLMs output and to keep lower the risk of hallucination. There are several techniques to design the prompt which help tailor outputs to the expected format and domain constraints. In a *zero-shot prompt* the model is instructed to provide a summary without providing examples. *Few-shot prompt*, on the other hand, presents a couple of examples of text and their summaries, and this has proven to dramatically improve the performance (Brown and al. 2020). Another technique is *Chain-of-Thought (CoT) prompting*, in which the model is prompted to reason step by step instead of asking directly to provide an answer. On complex tasks the model performs better using the CoT prompting, since it can encourage the model to internally structure the information before compressing it (Wei and al. 2022). Moreover, *role prompting* method is the technique in which the model is designated a specific role, which allows the outputs to be more tailored to the specific domain (Xu 2023). In summary, LLMs, with careful prompt engineering, are demonstrating to offer powerful summarization capabilities that can be exploited for pharmaceutical operations. They can automatically condense unstructured reports into digestible knowledge. The next challenge is ensuring these models have access to accurate, domain-specific information when generating summaries, which leads to the concept of retrieval augmented generation.

3.4 Retrieval-Augmented Generation (RAG)

In order to assess the problem of hallucination, besides prompt engineering researchers have developed a hybrid solution, to exploit both generation capabilities of LLMs and controlled knowledge base thanks to information retrieval techniques. This technology is called Retrieval-Augmented-Generation (RAG) which refers to architectures that combine an LLM's generative process with an external knowledge base to produce informed, factual outputs (Lewis and al. 2020). Lewis and al. (2020) introduced this method in 2020, and demonstrate augmenting generation with retrieved evidence can yield more accurate and up-to-date answers. This approach combines the strengths of pre-trained generative models, which store knowledge in their internal parameters (parametric memory), with external sources of information (non-parametric memory), such as document databases or retrieval systems (Lewis and al. 2020). By integrating both components through a general-purpose fine-tuning strategy, the model can generate more accurate and context-aware responses (Lewis and al. 2020).

Therefore, in the RAG system the LLM is combined with a retriever which, given the user query, pulls the top-k relevant documents from a corpus, and then the LLM generates a grounded answer that explicitly includes those documents. Lewis and al. (2020) RAG models used a large seq2seq model as the generator and Wikipedia as the external text corpus, showing strong performance and open-domain question answering with proper citations. The same principle applies to the summarization of shift handover reports: before generating a summary, the system retrieves the most relevant pieces of information and provides them to the language model as context. Based on this curated input, the model then produces a coherent and grounded summary. Indeed, the RAG approach is highly suitable for our knowledge management scenario.

The RAG framework is highly relevant to our knowledge management scenario. By design, RAG addresses two key needs: provenance, since there is the source documents of the output can be traced back, and dynamic knowledge update, since there is the possibility to handle new data added to the knowledge repository. In a

pharmaceutical manufacturing environment, these are key features to assure a higher probability of trust and, therefore, usability of a KM system.

3.5 Agentic AI and Orchestration Frameworks

In recent years, the increasing complexity of tasks delegated to AI systems has led to the emergence of agentic AI architectures. These systems are designed around the concept of autonomous agents that are capable of perceiving their environment, making decisions, invoking external tools, and coordinating multi-step workflows to achieve specific goals.

This approach marks a departure from monolithic AI pipelines toward modular and composable systems, where individual components, which are retrievers, language models, document processors, and human input handlers, can be orchestrated dynamically. Such flexibility is particularly valuable in knowledge-intensive scenarios, where multiple tools must be integrated to handle semi-structured data, user queries, and summarization.

There are increasing frameworks capable of supporting the design of agentic AI systems. Namely, AutoGPT (AutoGPT 2025), CrewAI (CrewAI 2025), LangChain (LangChain 2024) and LangGraph (LangChain AI 2024). In particular, LangGraph, which is a recent extension of LangChain, makes use of the structure of graphs to orchestrates the different tools of an agentic AI system. LangGraph is an open-source framework that makes it easier to create and manage complex workflows powered by large language models (LLMs). The AI system is modeled as a graph and each node is assigned to a task or tool, becoming specialized in that function. Nodes are connected by edges that define the workflow. There is a central node, which is the agent, that has full awareness of the environment and calls the tools needed to complete tasks. According to IBM's technical description, *LangGraph is designed to build, deploy and manage complex generative AI agent workflows* and uses graph architectures to manage relationships between components (IBM 2024). For complex decision making LangGraph allows the inclusion of conditional loops.

This architecture facilitates the design of robust, stateful workflows that require

branching logic, asynchronous operations, or progressive reasoning. Features such as streaming outputs, context-aware memory, and edge-based control flow make it well-suited to applications involving complex document processing pipelines and knowledge integration (IBM 2024). IBM notes that LangGraph's state feature acts as a "*memory bank that records and tracks valuable information processed by the AI system*", akin to a digital notebook of intermediate results IBM 2024 For example, if an agent has extracted text from a PDF, it can store that in state for a subsequent agent to summarize. This memory enables the system to tackle multi-step tasks where the outcome of one step influences the next, without losing track of context.

Another important aspect is the human-in-the-loop (HITL) capability. LangGraph and similar orchestration tools allow certain checkpoints where a human user or expert can review or correct the agent's output before it proceeds. This is particularly important in high-stakes environments like pharma: the AI might draft a summary or recommendation, but a human may need to validate it. The LangGraph framework explicitly supports HITL; as IBM describes, it *uses the human-in-the-loop approach* as part of its monitoring mechanism, meaning developers can design workflows that pause for human approval or input IBM 2024. This ensures the system remains under human supervision for critical decisions, combining automation with oversight

Agentic frameworks provide a suitable infrastructure for industrial knowledge management use cases, where tasks such as retrieval, filtering, summarization, and user interaction must be combined coherently. In such contexts, agentic orchestration enables the system to adapt to document heterogeneity, handle long-running operations, and offer guided exploration of results, improving both accuracy and usability.

3.6 Identified Gaps in Existing Solutions

Regardless of the discussed in the previous sections, the current knowledge management solutions in a pharmaceutical setting still present notable gaps.

First of all, a lack of domain-specific processing can be identified. Namely, the

fact that most generic knowledge management solution are not explicitly tailored and designed for the pharmaceutical production domain. Therefore, they might not understand some specific terminology, acronyms, or regulatory context. Vendor solutions like Eschbach's SAMI, demonstrate the importance of domain adaptation, given the tailored training for pharmaceutical industry to search and retrieve data effectively (Eschbach 2023).Existing enterprise search platforms are often too generic, and internal MES/DMS tools lack NLP capabilities (Eschbach 2023).

A second gap that can be identified refers to the missing integration of retrieval and summarization solutions. The concept of RAG is still a cutting edge solution and not widely used as an integrated industrial tool. Users are have to read many documents and procedures, and in a second moment synthesize information. This is labor intensive, since it requires much effort from the users, and is also prone to human error or oversight. There is a gap of an automated pipeline that brings together the query of the user, the retrieval and the synthesis, able to semantically capture the meaning of the content processed allowing for an effective and correct output.

Moreover, there is, still, high reliance on manual and email-based information sharing in many production environments today. Knowledge transfer happens much via informal channels and in a heterogeneous manner. There is a lack of structured and standardized solution for sharing and extracting relevant knowledge from the set of day-to-day information that flows in production settings. As Davenport (2025) pointed out, simply implementing tools like Microsoft 365 (SharePoint, etc.) doesn't automatically solve the knowledge problem – a lot of content ends up as unstructured files that people still have to search manually (Davenport 2025). The gap here is the absence of a central, intelligent knowledge base that employees can query conversationally instead of relying on asking colleagues or sifting through old emails. Moreover, important knowledge in those emails may not be captured elsewhere, leading to repeated mistakes or duplicated efforts.

Another subtle gap is how existing systems either rely purely on manual effort or, in the case of some AI solutions, operate as black boxes with no user control on the information processed. Bridging this gap would greatly enhance user trust and

the system's utility over time.

The technologies discussed in this chapter represent the state of the art in knowledge-intensive systems. However, their integration into a unified, context-sensitive solution for industrial KM remains largely unexplored. In chapter 4 the thesis shows the method followed in integrating all the technologies explored above into one single solution to the knowledge management problem stated in chapter 2.

Chapter 4

Methodology

This chapter aims at presenting the methodological choices taken to carry on the design of the proposed solution to the problem stated in Chapter 2. This is carried out thorough a description of the exploration of the documents, the pre-processing and cleaning of the data, and the development of a strategy to efficiently retrieve relevant information and elaborate it into useful and meaningful findings.

4.1 Data Exploration, Cleaning and Pre-processing

The first step of the whole process was all about understanding the right way to approach the given data. In Chapter 2, Section 2.5 the configuration of shift handovers has been widely described.

The documents are divided in different sections, some made of pre-populated fields, others are open so accept free text written by the user. Text is made both of numbers (total produced in batch, percentages of scraps, total produced in the shift, etc.), of unified codes (batch ID, product ID, work orders, etc.), machine and station names and other official or unofficial acronyms. Language is a mix of Italian and English. The writing style is fragmented and non-discursive, often consisting of short, bullet-like entries rather than full sentences. Moreover, differences exist between the two lines taken at the initial phase of development of this project, highlighting how, even within the same organization, different habits and patterns can emerge in the way this type of knowledge is managed. Designing a solution capable of handling

a wide variety of documents, flexible enough to accommodate customization, yet robust enough to extract the maximum amount of information, emerged early on as both a challenging and critical requirement shaping the project's methodology.

The data explored required a dense cleaning and pre-processing. To tackle the problem initially, static data was extracted from the organization's database covering the period from September 2024 to April 2025, pertaining to two different device and packaging production lines: Line A and Line B. It was decided to process a limited and static dataset in order to start in a simpler and controlled environment.

The entire textual content was stored within a single dataset entry. To facilitate efficient data handling, the following cleaning strategy was adopted:

- **Segmentation** of text into three distinct sections—Batch Information, Production, and Maintenance—for each sub-production line.
- **Removal** of entries corresponding to production shutdown days, which contributed noise without informative value.
- **Generation** of dedicated PDF documents for each line, thus providing clear reference material for consultation and troubleshooting during the development phase. During the development stage of the system the PDF documents to be processed arrived to 2260.

It is worth mentioning that a significant portion of the cleaning and pre-processing work was dedicated to handling the shift handovers coming from two different lines (Line A and Line B). Although both belonging to the same production module-device and packaging-they follow different approaches in filling these documents. These differences required separate pipelines; in fact, custom cleaning and pre-processing routines were implemented for each.

4.1.1 Metadata and Temporal Structuring

Based on the requirements identified from the context, a fundamental step in the development process was defined: keeping track of metadata for each document throughout the entire workflow. This ensures the traceability of the information provided to the end-user by clearly identifying its original source.

As mentioned above metadata such as date, shift name, and line name play a key role in document organization. They enable:

- Filtering and temporal segmentation
- Linking events across shifts
- Trend recognition and issue escalation tracking

Metadata were extracted from the original source filename using pattern-based information extraction. As a matter of fact each PDF document was accurately created posing attention also to the filename, which was already understood as a fundamental source of traceability. For example, filenames followed a structure such as: `"2025-01-15_Shift2_LineB.pdf"`. The extraction of metadata from filenames was performed through Python's built-in `re` library (Foundation 2024), leveraging regular expressions to identify relevant components and metadata embedded within the documents.

4.2 Document Augmented Retrieval

Subsequently, a document retrieval strategy was developed. Given the semi-structured and heterogeneous nature of the data, a hybrid approach was chosen. The system combines:

- **Semantic search**: using OpenAI's `text-embedding-ada-002` (OpenAI 2022b) model and `Chroma` vector store (Chroma 2024).
- **Keyword search**: via Langchain's `BM25` (LangChain 2024) to ensure keyword matching.

The reasons for this choice are explained below.

4.2.1 Semantic Retrieval with Chroma

To achieve an effective and valuable solution, it was crucial to leverage the semantic nuances inherent in textual data. This goal could be accomplished through semantic

retrieval methods, specifically utilizing text embedding techniques. Text embeddings are numerical vector representations capable of capturing the semantic essence and contextual relationships within textual information (Harsoor 2024). By mapping textual data into vector spaces, embedding methods facilitate semantic search, enabling the retrieval of documents based on conceptual similarity rather than simple keyword matching.

The dedicated vector store solution chosen for this project was `Chroma`, because of its inherent ability to deal with metadata filtering when performing similarity search (Amjad 2024). Employing such an embedding-driven semantic search mechanism allows for superior retrieval quality, effectively addressing challenges posed by heterogeneous textual data and thus significantly improving knowledge discovery and management.

4.2.2 Keyword matching with BM25

Semantic search alone was not enough, as it lacked the ability to capture highly specific situations. For instance, cases in which users could be interested in looking for problems encountered with a particular batch or machine on a specific day or within a defined period, cannot be efficiently retrieved using embedding-driven semantic search alone.

This limitation was addressed by incorporating sparse embeddings, commonly known as keyword-based search using the `BM25` algorithm. `BM25` evaluates the frequency of search terms within documents while also considering document length, thus enhancing the retrieval of highly targeted and relevant results (Desai 2023).

4.2.3 Hybrid Document Retrieval

A hybrid search solution combining semantic and keyword-based methods was selected to leverage the complementary strengths of both approaches. Each document is represented both as a semantic vector and via token frequency, enabling efficient retrieval even when queries vary in terminology. Semantic search effectively collects context and finds conceptually related information but struggles with precise queries, such as those involving specific batch numbers, machine identifiers, or exact

dates. On the other hand, keyword-based methods like `BM25` are highly efficient at retrieving precise matches by analyzing term frequency and document length, yet may overlook contextually similar content. Integrating these methods into a hybrid strategy enables robust and accurate retrieval, effectively addressing both broad semantic searches and highly specific queries, which is critical for efficient knowledge management in production environments.

From a practical perspective, the scores obtained through semantic and keyword-based methods were combined by assigning equal weights to each, ensuring both retrieval approaches influence the final ranking in the same way. Particularly, the retrieval process first employs `Chroma` to perform semantic similarity searches, yielding a set of relevant documents along with their scores. Concurrently, the `BM25` algorithm retrieves documents based on keyword matching. To effectively merge these results, duplicates are removed, and the remaining documents retrieved with keyword search are assigned cosine similarity scores calculated between their embeddings and the query embedding. The combined list is then re-ranked based on these unified cosine similarity scores, ensuring that both semantic context and precise keyword relevance contribute equally. The output of this hybrid retrieval method is thus a ranked list containing the top k most relevant documents, where k is a predefined parameter.

4.3 Summarization Strategy

After having identified a retrieval strategy the following step of the process included developing a component of the system capable of transforming the natural language content of the retrieved documents into useful insights through a structured summary. To perform this task, it has been selected the application of state-of-the-art technology, namely, Large Language Models (LLMs). For this project OpenAI's `GPT-4o` (OpenAI 2024) was deployed.

An important detail included in the deployment was the parameter *temperature* of the model, which was set to zero. This choice was made to ensure maximum determinism and consistency in the generated summaries, minimizing randomness in the output. In the context of summarizing shift handover reports—where factual

accuracy and reproducibility are essential—this setting helps to avoid hallucinations and guarantees that the same input will always produce the same output.

4.3.1 Prompt Design

The model was accessed through OpenAI's API using a tailored prompt designed in a system+user format. The prompt has been formulated combining role-based conditioning and instruction-based zero-shot prompting, enriched with contextual information extracted from the retrieved documents. Specifically, the system message defined the role of the model as a technical assistant, providing it with domain-specific knowledge of the production lines and corresponding machine names, as well as clear behavioral instructions (e.g., to always extract metadata such as shift and date, and to return information only when relevant, to avoid hallucinations). It was in this way that the model's responses were aligned, topic-related, and domain specific in terms of requirements. The user message contained the actual task asked by the user and the full text to analyze. What the model produced was a structured summary, broken down by significant sections (such as production events or machine-related issues), and always included the corresponding metadata. This structure made the extracted insights easy to interpret and useful for downstream analysis or user review.

This kind of prompting plays a crucial role in making the system highly tailored to the specific needs of the use case, indeed, offering a competitive advantage over existing off-the-shelf solutions. Commercial tools, while powerful, are often designed for general-purpose applications and lack the flexibility to effectively address domain-specific requirements such as those encountered in this context. The use of custom prompts allows the system to operate with a deeper contextual understanding, incorporating domain knowledge and behavior conditioning that would otherwise require less adaptable alternatives.

4.4 Consolidation into an Agentic Architecture

At this point the components described above had to be combined in a single solution giving to the users the possibility of interacting with it in the most seamless and

smooth manner possible.

To develop this idea it was decided to orchestrate an agentic architecture using `LangChain`(LangChain 2024) framework: more specifically `LangGraph` (LangChain 2024) library.

The reason behind the decision of developing an agentic architecture to this knowledge management problem stands in the need of a dynamic solution able to deal with a semi-structured context: an orchestrated and modular approach allows flexibility, reusability and a control over the information flow (LangChain AI 2024). `LangGraph` library was chosen because it is a widely used solution that allows to represent explicitly the agent's information flow, modular steps and handling the graph's states (LangChain AI 2024). It is able to orchestrate complex scenarios with also conditional steps (LangChain AI 2024).

4.4.1 Description of the Architecture

Figure 4.1 illustrates the architecture of the Agentic AI system developed for this project. The diagram visualizes the modular and orchestrated flow implemented through `LangGraph`. This structure captures the sequence and interaction between key components that enable query resolution through retrieval and summarization mechanisms.

The architecture consists of the following main components:

- **User query**: the user provides the input composed of a natural language query, a date range, and a selected production line. The user interacts only with the Agent.

- **Agent**: orchestrates the overall workflow, managing the transitions between tools based on the current state and the data provided. It does this in a completely autonomous way, following the instructions o each tool and interpreting the user query.

- **Document retrieval tool**: a hybrid search system (semantic similarity and keyword-based retrieval), as described in Section 4.2, which retrieves relevant shift handover reports from both Line A and Line B based on the user's query.

The output is a list of documents showing the filenames which include the metadata.

- **User selection tool**: tool that allows the user to review the list of documents given by the first tool and and select the documents deemed most relevant. The user can also select all the documents from the set if considered necessary. However, it is still the Agent which orchestrates the interaction with this tool, there is no direct communication between this and the user.

- **Structured summary tool**: generates structured summaries of the selected documents, tailored for each production line. The summarization logic and formatting strategy are detailed in Section 4.3.

This modularity provides high flexibility, allowing each component to evolve independently. It also enhances interpretability and facilitates debugging, as each step of the process is clearly separated and traceable. The architecture is designed to support extensibility, making it possible to integrate additional components such as the inclusion of other production lines.

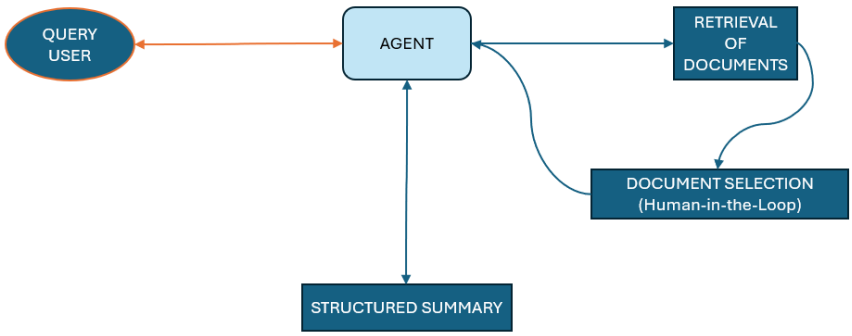

Figure 4.1: Architecture of the agent-based system built using LangGraph. (Own elaboration)

4.4.2 Main Features

Modularity

The system's architecture was deliberately designed to be based on a modular structure in such a way that each component—document retrieval or summarization—is implemented as an independent tool. Notably, dedicated implementations of these tools were written for each production line (Line A and Line B) taking into account the different ways in which documents are completed and structured in the different shifts in each line. Thanks to modularization, each tool can be refined or replaced independently and with a specific focus without having to adjust the system's other components. This makes it easier to maintain, iterate, and be robust. In addition, this separation of concerns allows for better scalability.

State Management within the Graph

The architecture uses persistent state management to pass and update information across the graph's nodes. This state includes all relevant elements of the interaction, such as the user's query, selected dates, production line, and the retrieved documents. Keeping each step in a consistent state allows the system to provide coherence in conversations but, most importantly, allows for scalability towards more advanced applications in the future. This is also made easier in terms of better traceability and modularity as each component operates within a clear and shared context.

Advantages over Monolithic LLM approaches

Compared to monolithic language model agents—where the entire process is handled in a single chain, the modular and graph-based setup offers several critical advantages. First, it allows component-level control over each step of the workflow, improving traceability and transparency. Second, debugging becomes significantly easier, as each node in the graph can be isolated and tested independently. The system is designed to support the replacement of individual modules: for instance, a different summarization model or retrieval engine can be integrated without affecting the rest of the system. Lastly, it helps avoid the black-box behavior often associated

with single-prompt solutions, fostering better understanding and control over the logic of each decision.

4.5 User Interaction

A crucial instrument of traceability and accountability was the inclusion of a *user interaction* solution within the architecture. It is the Agent which communicates with the user: once the query is issued and the document retrieval tool has been called and gave as a result a list of the top-k shift handovers, together with their metadata, these are shown to the user who can select which of the documents he or she is interested in receiving a structured summary. The user can also stop here and read the original documents, this is possible thanks to the printed metadata. This active role enables personalized and context-aware analysis while minimizing cognitive overload. Moreover, by involving the user in the document selection step, the architecture increases the overall reliability of the system, reducing the risk of irrelevant or hallucinated outputs generated by the summarization component.

4.6 Summary

This chapter presents the whole methodological steps that have been implemented to deploy the final solution for the problem stated. It starts from the first phase of exploration and preprocessing ending with the integration of all tools in a unique Agentic Artificial Intelligence system: each step of the process served as a foundation for the subsequent decision, contributing to the transformation of unstructured operational knowledge present in a great amount of documents into accessible and practical insights.

Chapter 5

Results

This chapter aims at presenting some use case results of the digital solution proposed to tackle the problem stated in Chapter 2. It presents all use cases which test the solution on different aspects. It shows the result of the system to some chosen queries and interprets the result.

5.1 Use Cases

5.1.1 Use Case 1: Frequent Mechanical Issues on Line A in a given month

The first use case brought involves a high-level question in which the user asks which were the most frequent mechanical problems and how they have been addressed in Line A in a given month (February 2025). The following table 5.1 summarizes key findings:

Category	Shift	Issue Description	Resolution	Area
Box jams at Machine 1	23/02 (Morning)	Box and carton jam causing stoppage from 12:26 to 12:33	Removed cartons, machine reset and restarted	Mech.-Electrical
Machine 2 issues	23/02 (Morning)	Lot launch failure from 06:00 to 06:35	Support called, retry attempts on machine	Mech.-Electrical
	25/02 (Afternoon)	Flag-like label application and pen jams behind drum	Changed label reels until good one was found	Mech.-Electrical
Pen jams at Machine 2	02/02 (Morning)	Repeated slowdowns due to glue-rich labels	Unresolved	Mechanical
	02/02 (Morning)	Multiple pen jams, total 25 minutes downtime	Jams removed, reset and restart	Mech.-Electrical
Cap feeder issues	25/02 (Afternoon)	jam in cap feeder 1 delivery	Not specified	Mech.-Electrical
	02/02 (Morning)	50 min stop, missing nut and detached paddle	See work order XXXXX	Mech.-Electrical
Machine 3 problems	28/02 (Night)	Product overload alarm from 00:59 to 01:09	Checked by maintenance team	Mech.-Electrical
	28/02 (Night)	Scraps due to variable data not read	Parameter changed, see work order XXXXXX	Mech.-Electrical

Table 5.1: Summary of mechanical problems in Line A - February 2025

The system successfully aggregated relevant events, showing, for example, repeated issues with the labeler due to label quality (excessive glue), and frequent jams in the cap feeder. These patterns could inform future preventive actions or procurement adjustments.

The first use case highlights several key capabilities of the proposed solution. Primarily, the system demonstrates targeted and combined retrieval, integrating both temporal filtering (e.g., retrieving documents from February 2025) and context-specific constraints (limiting the scope to a particular production line). A major strength of the system lies in its ability to identify semantically relevant events, even when expressed in varying terms, such as recognizing the equivalence between phrases like *pen jams*, *machine stoppages*, or *slowdowns*.

In addition, the generated response is not a flat list of events but rather a structured and categorized summary. Problems are grouped by type or equipment, and recurring patterns or trends are made explicit. Each entry includes key information such as the area involved and the corrective actions taken.

5.1.2 Use Case 2: Comparison between 2 shifts of the same day (10 March 2025) in line A

The following table 5.2 displays the answer of the system to Use Case 2. In this case, the question was asking what happened during the night shift with respect to the day shift of a specific date in line A, for a shared production batch, highlighting differences in output and production-related events.

Aspect	**Night Shift**	**Day Shift**
Shift Report	2025-03-10_ShiftNotte.pdf	2025-03-10_ShiftPomeriggio.pdf
Units Produced	10,600	1,800
Main Issues	Downtime from the beginning of the shift until 23:10 due to pen replenishment, following the detection of a critical defect carried over from the previous shift. This was followed by unpacking and a 200% visual inspection of one of the pallets affected by the identified defect. A further 30-minute stoppage occurred due to an alarm: "Robot 1: products are located in the insertion area," which was resolved by contacting maintenance (ODL XXXXXX).	Downtime from 14:00 to 18:45 due to an investigation conducted by the Process Technician following the detection of a critical defect. The issue was resolved with ODL XXXXXX. Additionally, an adjustment was made to labeler head 2 (ODL XXXXX).
Defect Origin	Carried over from previous shift	Same defect, root cause analyzed
Maintenance Involvement	Active resolution of two issues	Single long investigation

Table 5.2: Comparison between night and day shifts on Line A – 10 March 2025

The second use case further illustrates the system's ability to extract, compare, and contextualize production data across different shifts working on the same batch. By aggregating relevant events from both the night and day shifts of 10 March 2025 on Line A, the system highlighted a significant disparity in productivity (10,600 vs. 1,800 units), despite both shifts being affected by the same underlying defect.

This case demonstrates the solution's strength in performing comparative diagnostics, surfacing discrepancies in operational responsiveness and the effectiveness of corrective actions. The system provides not only numerical output comparisons, but also contextual insights, such as the nature of the downtime, the involvement of maintenance teams, and the specific interventions applied in each shift (e.g., pen replenishment and visual inspections at night vs. extended defect investigation and

mechanical adjustments during the afternoon).

In doing so, the system enables cross-shift alignment and continuous improvement. These comparative insights allow stakeholders to identify more efficient practices and standardize resolution protocols across the organization.

Additionally, this use case reinforces the value of having a centralized, searchable knowledge layer, where shift-level events are not only preserved but can also be systematically analyzed and compared. This resulted in being a key enabler for operational transparency and process optimization.

5.1.3 Use Case 3: Asking for a specific batch in a given month

The third Use Case tests the solution's ability to look for a specific batch ID but in a generic month. The system retrieves a list of seven documents as it is programmed to do. Those documents refer for the most part to the days in which that batch has been produced, however there are also other days: this because it is required to find the best k documents, the user still can select which documents he or she is more interested into summarizing. Table 5.3 displays the result from the system. Given the scope of this work, only a relevant excerpt of the output is shown, as it sufficiently demonstrates the system's behavior and the type of information retrieved.

Shift	**Production**	**Issues**
Morning, 2024-10-03	In shift: 23112, Total produced: 61800, Total for batch: 62622	*Machine 1*: Numerous alarms "monitoring parts in separation present" in station x. *Machine 2*: Slowdowns due to "Missing label" on xxxx 2. Downtime from 12:00 until the end of the shift due to server patching activities.
Afternoon, 2024-10-03	In shift: 10032, Total produced: 62520, Total for batch: 62622	*Machine 5*: Multiple stoppages due to carton jams at the conveyor exit *Machine 4*: 33-minute stop caused by incorrect label adhesion on shipping boxes, required mechanical checks, multiple stoppages due to incorrect pickup of shipping boxes, resulting in slowdowns for product replenishment, mechanical checks required.
Afternoon, 2024-10-02	In shift: 8708, Total produced: 8708, Total for batch: 62622	*Machine 1*: 10-minute stop following a crash at station x *Machine 3*: Downtime from the beginning of the shift until 14:15 for screw replacement on pre-stopper fork 2/2.2, mechanical intervention required, numerous rejects, mechanical checks in progress. *Machine 4*: Multiple stoppages due to label adhesion failure on shipping boxes, causing slowdowns for product replenishment, replacement of tape blades, mechanical intervention required

Table 5.3: Information on a specific Batch

This table, which only displays a part of the output, shows the ability of the system to provide correct and precise answers even if the user requests for a specific batch in a generic month. This is an example of how relevant is the implementation of hybrid search retrieval, which both addresses cases in which a keyword match is required and semantic understanding of the request.

Finally, this case highlights the system's potential in batch-level traceability. It allows users to reconstruct a coherent production timeline by aggregating fragmented information across shifts and days. This feature is especially valuable in contexts requiring quality audits, investigations, or compliance reporting in regulated environments such as pharmaceutical manufacturing.

5.1.4 Use Case 4: Frequent alarms on a specific machine in a month

The fourth Use Case worth mentioning investigates whether frequent alarms occurred on a specific machine of a production line in a given month are to be brought to the attention of the user because they may show a bigger problem to be addressed. The system retrieved and summarized multiple shift reports, highlighting repeated alarms across several dates and shifts, which could indicate underlying mechanical or control system issues. Results are displayed below in table 5.4.

Date and Shift	**Reported Alarms and Notes**
26 Feb 2025 (Afternoon)	Frequent slowdowns due to *overload detected* alarms on all tracks at St.18 and St.30/31. Continuous slowdowns at St.18 due to *waiting part* alarm on track 3. Mechanical intervention performed (ODL xxxxxx).
14 Feb 2025 (Morning)	Frequent *overload detected* alarms at St.18 and St.30/31. Frequent *waiting part in separator* alarms at St.18.
16 Feb 2025 (Afternoon)	Frequent *overload detected* alarms at St.08 (track 2) and St.18 (tracks 3 and 4).
17 Feb 2025 (Night)	Frequent *overload detected* alarms at St.08 (track 2), St.18 (tracks 1 and 4), and St.30/31.
5 Feb 2025 (Night)	Multiple *overload detected* alarms at St.08 (tracks 1 and 4), St.18, and St.30/31. Multiple *waiting part in separator* alarms at St.18 due to slow RNS movement along linear guides. Mechanical optimization required (ODL xxxxxx).
8 Feb 2025 (Night)	Frequent jamming of the RNS Puller in the BF along the linear guides at St.18, on all tracks.
These recurring alarms may indicate systemic issues that require further investigation and interventions in order to improve the operational efficiency of machine 1.	

Table 5.4: System-generated report of recurring alarms on Machine 1 – February 2025

This case highlights the system's capacity to detect and synthesize recurring low-level signals that may otherwise be overlooked in daily operations. Despite the user's vague query ("*Are there frequent alarms on Machine 1?*"), the system correctly filtered and aggregated relevant reports, showing consistent patterns across stations and dates. Additionally, by linking each alarm with its corresponding work order or mechanical intervention, the system provides insightful data for reliability engineering and continuous improvement efforts.

This use case demonstrates the value of the system not only in retrieving facts, but also in surface-level signal mining, allowing for proactive maintenance planning, early anomaly detection, and operational diagnostics, even when the user query is incomplete or exploratory and general.

5.1.5 Summary of Capabilities Demonstrated

The four use cases presented in this section provide a comprehensive overview of the core capabilities of the proposed system and its applicability to real-world production

scenarios in the context of pharmaceutical manufacturing.

First of all the system is capable of retrieving relevant documents by an efficient combination of temporal filters and contextual requirements, such has production line or mentioning a specific machine or station. Second, the system does not simply list the information contained in the documents but generates a structured summary grouping issues, making some patterns explicit. Use Case 2 shows how the system can also compare multiple documents extracting insights on the same issue across shifts. On the other hand, Use Case 3 shows that the system can reconstruct the evolution of a unique batch over multiple shifts, even if in the query the exact period is not expressed. Whereas, Use Case 4 demonstrates the ability to identify systemic issues based on the repetition of seemingly minor alarms across different shifts.

All cases show how the system can be a useful support in decision making without never substituting the user. This last, indeed, has the possibility to select which document to inspect among those that the system proposes. Moreover he or she can always go to check the original source thanks to the document's metadata which are always displayed together with the insights. Final selection and interpretation are left to the user: fundamental feature for traceability and trust in the system by pharmaceutical manufacturing professionals.

Overall, the system reveals its versatility in supporting various tasks.The results underline its potential as an intelligent interface between operational knowledge and engineering decision-making.

Chapter 6

Evaluation and Discussion

This chapter goes through the all evaluation of the solution proposed in this work and widely explained up to now. The chapter first shows the evaluation criteria chosen to measure such a complex system made of different modules, and developed in a particular context for which the engagement of domain experts resulted to be a fundamental contribution. It follows by displaying the results of such evaluation, followed by an interpretation of these results, identifying the criticism of such system, also compared to a commercial general-purpose solution. In conclusion, some further developments are described, as crucial innovations to reach greater significance and trust in such system in a pharmaceutical manufacturing environment.

6.1 Evaluation Criteria

Assessing the performance of a custom-based and complex system as this one is particularly difficult. Especially given the domain for which it has been designed, the performance has been evaluated giving a special attention to the reliability of its outputs.

Choosing a qualitative human evaluation was judged as the only way to critically evaluate the performance of the system developed. The challenge presented by the organization referred to the possibility of this custom based system not being enough better performing that another commercial and general-purpose solution off-the-shelf which has been already integrated by the company in all its systems. The

off-the-shelf model in question is Microsoft Copilot, a black-box general-purpose solution purchased by the company and safeguarded by software that enables users to include confidential and proprietary information in their queries. Since the model was not developed in-house but rather acquired commercially, it remains a black box. Consequently, it is not possible to investigate how it processes data or how it makes decisions.

The most straightforward but reliable method chosen was the generation of a set of hypothetical questions which may cover the most variability. Testing the custom system's ability to respond and comparing results to those coming from the solution present on the shelf. The results of both solutions were given to a group of potential users of the system expert in the field.

The results were evaluated by field experts which qualitatively and subjectively assigned a score from one to five for three different criteria: Answer Relevance, Faithfulness and Coverage. These are commonly adopted qualitative metrics in the evaluation of advanced retrieval-augmented-generation (RAG) systems (Harchaoui 2023). Nevertheless, it must be underlined that the subjectivity of the group of evaluators must be taken into consideration. Each expert assigned a score from one to five for each metric based on their professional judgment. Whenever the system is not able to answer, the score assigned is equal to one.

The metrics chosen where:

- *Answer Relevance*: which tests how good the results answers to the question asked.
- *Faithfulness*: which tests whether the answer is reliable and no hallucinations are made.
- *Coverage*: which tests whether all the important aspects relevant to the questions and present in the PDF documents are covered.

6.2 Domain Experts Evaluation

Integrating human judgment in the evaluation process presents both advantages and disadvantages.

Among the advantages it must be recognized the context for which this system has to be deployed. Human judgment, indeed, provides a validation that this system interprets correctly details specific to the domain in question, being critical for regulatory audits. Human expertise helps to counterbalance the dangers of wrong or partially accurate Artificial Intelligence outputs that may lead the user to wrong conclusion that can create serious operational, quality or even safety problems. Probably the most straightforward advantage is that it may represent an instrument to gain end-user confidence in the system. Within a pharmaceutical environment building trust in a system is an essential ingredient in generating adoption of the system itself. Especially because this system has to be integrated with well established and structured scenarios. The potential users are surely more likely to trust a system that has been stringently validated across their peers or well-trained domain experts. Lastly, including human judgment and evaluation supports continuous improvement, identifying scenarios where the custom-developed solution outperforms existing commercial solutions, thus justifying investment in custom tools.

On the other hand including human evaluation presents some obstacles and disadvantages. First and foremost engaging human experts increases, inevitably, time and costs linked to the evaluation process of the system, especially considering the necessity of including resources highly qualified in the pharmaceutical manufacturing sector. This also means that it takes operational staff away from their core field activities, potentially impacting production continuity and efficiency. Moreover, in contrast with the rapidity in innovation and integration that may be expected by these kind of systems, implementing and validating this solution in this way may have severe implication on the speed of deployment. Another critical perspective may underline that including human-in-the-loop techniques brings an intrinsic subjectivity in the outcome of the evaluation, and it results difficult to standardize judgment criteria on qualitative metrics like *faithfulness*, *relevance* or *coverage*. In

addition, the sectorial expertise, even if valuable, can bring human expert to end up in cognitive biases and prejudices: so they may tend to confirm information already known, and over or under estimate some risks based on personal experiences. Finally, a further disadvantage may reside in the limited scalability of a human evaluation technique. As mentioned in 2.5.3, the configuration of these documents makes their volume to grow very quickly. This rapid scaling up of documents may require validating much more documents, therefore, this evaluation criterion can end up being a serious operational bottleneck in the deployment.

6.3 Test Queries

In order to generate test queries in a structured and representative manner, a Large Language Model (LLM) was used. The model was provided with clear instructions requiring it to process PDF documents and produce a diverse set of questions, varying significantly in specificity. Some questions were notably specific, focusing on information related to particular batch IDs, maintenance events, alarm occurrences, or incidents during a defined shift or date. On the other hand, other questions were more general, addressing broader scenarios such as identifying shifts with the highest down time or percentage of scraps over extended periods, pinpointing recurring alarm and associated down times for specific machines, or highlighting notable production issues occurring within a certain month. Additionally, the questions covered the identification of personnel management problems, comparisons between production shifts within the same day or over a given time frame, and issues related to particular equipment or processes. The LLM was instructed to adopt the perspective of a hypothetical user persona, as the ones described in 2.4, familiar with the production environment, using appropriate acronyms and formulating concise, straightforward queries.

Initially, the model generated a total of 40 hypothetical questions, including an acceptable amount of duplicates or questions referencing batch IDs not present within the specified dates. This initial set was subsequently manually reviewed, revised, and enriched through collaboration with field experts, culminating in a final

evaluation set of 60 diverse and comprehensive questions: 22 more specific, and 28 more generally oriented.

6.4 Results

The test queries described in 6.3 were administered to both the custom-developed solution and the off-the-shelf commercial system already integrated within the company's services. The outputs produced by each system were reviewed by a pool of domain experts, who assigned qualitative labels according to the criteria previously outlined. The table 6.1 below presents the summary statistics of the scores assigned, offering a comparative overview of the two systems' performance across the selected evaluation metrics explained in 6.1.

Metric	**Custom Mean**	**Custom Std**	**General Mean**	**General Std**
RELEVANCE	4.28	1.03	2.42	1.80
FAITHFULNESS	4.70	0.79	2.43	1.81
COVERAGE	3.90	1.11	2.04	1.44

Table 6.1: Comparison of average scores and standard deviations across evaluation metrics.

The custom-based system obtains higher mean values in all three metrics, highlighting an overall more positive evaluation by experts. Differences in mean values are more pronounced for *faithfulness* and *relevance* metrics, pointing out more accurate and contextually appropriate responses. Another insight to highlight refers to standard deviation values: the custom system has lower standard deviation, which implies higher coherence outputs' quality, with respect to the standard deviation values of the off-the-shelf solution, which suggests inconsistent performance, with responses that are occasionally well-received but often fall short of expectations.

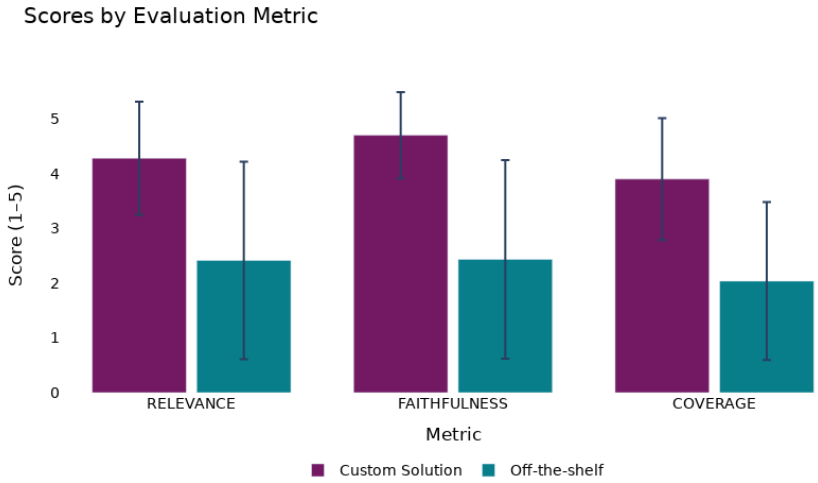

Figure 6.1: Comparison of average evaluation scores between the Custom Solution and the Off-the-Shelf system.

Figure 6.1 provides a visual comparison of the average scores per each metric. While it represents the same data shown in Table 6.1, the figure enhances interpretability by allowing for a more immediate understanding of the differences and trends discussed above.

By aggregating the scores across all evaluation metrics, a clearer picture of the systems' overall performance emerges. The Custom Solution achieved an average score of approximately 4.28, with a relatively low standard deviation of 0.64, indicating not only a high level of performance but also a strong consistency in the quality of its responses as perceived by domain experts. Conversely, the Off-the-Shelf system obtained a significantly lower mean score of 2.30, coupled with a higher standard deviation of 1.76. This reflects not only a generally weaker performance, but also a marked inconsistency in the results, with expert evaluations varying considerably depending on the specific query.

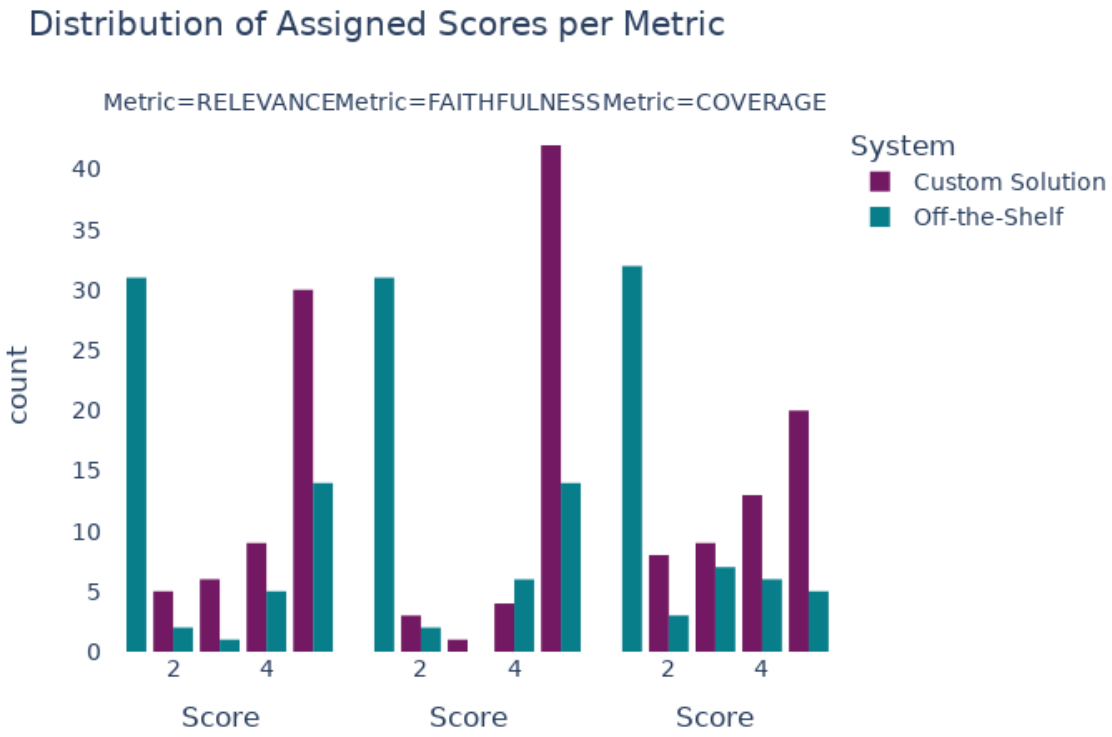

Figure 6.2: Distribution of expert-assigned scores for each evaluation metric, comparing the Custom Solution and the Off-the-Shelf system.

An interesting analysis to be included in the presentation of the results is the distribution of assigned scores per metric, displayed in Figure 6.2. The histogram presents several insights. For what concerns the metric *relevance*:

- the Off-the-Shelf solution shows a polarized distribution, with a strong concentration of scores equal to 1, and a secondary peak in the highest score (5). This points out that the quality of the response lacks of consistency: in some cases are considered not inherent to the questions, in other rarely optimal.
- On the other and, the Custom Solution shows a progressive increase in scores from 1 to 5, with a prevalent concentration towards high scores, suggesting a higher reliability and coherence in the capacity of generating relevant responses.

For what concerns the metric *faithfulness*:

- Also in this scenario the Off-the-Shelf is distributed at the extremes, with many scores equal to 1 and a relevant number of 5: an indication of instability in

generating source-grounded responses, potentially prone to hallucinations or serious inaccuracies.

- The Custom Solution has a highly unbalanced distribution towards the score equal to 5, with few lower scores, indicating that its responses are perceived as constantly reliable and accurate.

For what concerns the metric *coverage*:

- In this metric the distribution is much more balanced, but it must be highlighted that the Off-the-Shelf solution receives lower scores (1-2), while the Custom Solution has a more homogeneous distribution, slightly increasing towards higher scores, suggesting a more complete and constant information coverage.

Even if the systems were tested on a limited number of evaluation prompts, the perception and opinion of people actively working in this context has to be valued and it allows to draw some useful considerations.

The Custom Solution is perceived as more relevant, reliable, and complete, with a score distribution that reflects consistency among expert judgments. Scores tend to progressively cluster toward the highest values, and variability remains limited.

In contrast, the Off-the-Shelf solution shows significant weaknesses in terms of stability and quality. The high concentration of scores equal to 1 across all metrics is a critical signal, indicating that the responses were inadequate or misleading in a substantial number of cases. The significantly higher standard deviation observed for the off-the-shelf system further highlights its inconsistent and unreliable performance.

6.5 Final Considerations and Discussion

On the basis of what has been presented in section 6.4, some important considerations can be concluded. This must be done having always clear what is considered to be an essential requirement from the perspective of a professional operating in an industrial and highly regulated scenario. In such environment, reliability, traceability, and transparency are non negotiable prerequisite that information that passes thorough

production context should preserve. Building trust in users is one the most difficult and important objective that innovative solution as an Artificial Intelligent system have to reach in scenarios as this one in order to be successfully implemented and spread.

In this perspective the concept of Human-Centric AI has to be introduced, as an essential framework of inspiration for the development of the solution brought by this project. AI can bring disruption, which can have positive and negative effects at the same time. On one hand it is a tool for innovation and a boost of efficiency and speed in whatever process it is deployed, but on the other hand, it may present many challenges, given its complexity, novelty, and pervasiveness.

Schmager, Pappas, and Vassilakopoulou (2025) provides a systematic literature review of what is encompassed behind the concept of Human-Centric AI (HCAI) as an approach to AI design. The authors underline that *"a crucial first step is for AI designers, developers, and policymakers to begin envisioning how AI systems can be shaped to support human welfare, enhance human well-being and augment - rather than diminish – what it means to be human"* (Schmager, Pappas, and Vassilakopoulou 2025). This framework comes from the concept of Human Centric Design which implies an innovative method for problem solving that starts by a deep understanding of the target audience and its viewpoints, and it outputs a design which impacts positively the whole stakeholders involved.

Schmager, Pappas, and Vassilakopoulou (2025) review the possible defining elements of HCAI: its purposes, values and properties. It is the result of a change of paradigm in which rather than humans having to adapt to technology, suitability for humans must be considered from the very beginning of technology's design.

HCAI's purposes can be formalized in automation, augmentation and AI autonomy. The first denotes the undertaking of tasks by AI without human intervention, the second refers to the amplification and empowerment of human capabilities, skills and decision making possible thanks to AI, and the third points to the capacity of AI to act in an autonomous way performing actions (Schmager, Pappas, and Vassilakopoulou 2025).

Values of a HCAI, on the other hand, comprehend ethics, protection and per-

formance. Ethical AI includes dignity, justice and fairness in the whole process of design, development, deployment and use of AI solutions (Schmager, Pappas, and Vassilakopoulou 2025). Protection includes everything that deals with privacy, prevention of harm and safety (Schmager, Pappas, and Vassilakopoulou 2025). While the value of performance may cover three other aspects: an efficient technology that reaches the desired objective with the minimal resources, an effective technology that accomplishes outcomes in line with the goals identified by the context, and an accurate technology which generates reliable outputs (Schmager, Pappas, and Vassilakopoulou 2025).

The last dimension of HCAI solution highlighted by Schmager, Pappas, and Vassilakopoulou (2025) relates to the assurance of AI properties. These incorporate oversight-related aspects such as accountability, responsibility, and controllability, which ensure meaningful human supervision. Comprehensibility is equally important, as systems must be interpretable and understandable to users, stakeholders, and regulators. Transparency and traceability further support this by clarifying data usage and algorithmic processes, fostering trust and mitigating bias. Additionally, sustainability, environmental, social, and economic, is emphasized to ensure long-term viability. Finally, reliability and robustness are essential to guarantee consistent and safe performance across varied contexts.

This design framework is, undoubtedly, applicable to the pharmaceutical manufacturing context which constitutes the target domain of the knowledge management solution proposed in this thesis. It has served as a source of inspiration. Throughout the development of the custom Agentic AI system, careful consideration was consistently given to the specific characteristics of the pharmaceutical manufacturing environment, the potential end-users, and the challenges professionals face in managing knowledge derived from shift handover reports.

From an HCAI perspective, the custom solution developed in this work exhibits alignment with several key properties. It supports a degree of traceability, particularly regarding the origin of the data used for information extraction. Moreover, it incorporates human-in-the-loop mechanism: once the system retrieves a list of potentially relevant documents, the user remains in control by selecting which ones

should be passed to the summarization tool. This preserves human oversight in the knowledge extraction pipeline and aligns with the principle of meaningful human control.

Despite its superior performance in terms of output quality, the custom solution also presents a number of challenges and limitations that must be considered, particularly in terms of explainability, implementation effort, and resource requirements.

The system cannot be considered fully explainable. While users can inspect which documents were retrieved and which were summarized, they cannot access or understand the internal logic that determined why certain documents were prioritized or excluded. As a result, transparency and explainability remain partial, an this limitation may directly impact trust. First, while the system is partially traceable in its retrieval and summarization processes, it cannot be considered fully explainable or transparent from a business user's perspective. The internal logic, based on large language models and complex retrieval pipelines, is still largely opaque, especially for non-technical stakeholders, and this can hinder full adoption and trust. Trust in this system derives mainly from its observed superior performance, particularly in terms of faithfulness, when compared to an off-the-shelf solution as displayed in section 6.4. Nevertheless, this trust is empirical rather than structural: users may trust the solution since it provides good results and performance satisfaction, however they have no full picture and understanding of how the system works internally and the underlying decision-making process remains largely opaque.

Furthermore, coverage occasionally suffers: certain relevant details may be missing from the summarized output because some documents were not retrieved. This again ties back to the lack of control over deeper aspects of the retrieval process, reinforcing the need for more transparent and adaptable mechanisms.

From an implementation perspective, deploying such a custom system poses technical and organizational difficulties. Cloud-based implementations may conflict with company policies related to data security, governance, and regulatory compliance. On-premise deployment, while potentially more compliant, introduces additional complexity and cost in terms of infrastructure, maintenance, and operational stability.

Moreover, building and maintaining a custom solution requires a significant investment of time, financial resources, and specialized human capital. It demands not only technical development but also continuous validation, updates, and alignment with evolving business needs. This can represent a substantial burden for organizations that lack internal AI expertise or the capacity to sustain long-term technical projects.

On the other hand, the off-the-shelf solution is designed as a general-purpose tool, the same for all users and contexts. As such, it requires humans to adapt to the technology, rather than the technology being tailored to the specific needs, workflows, and constraints of its users. It offers an accessible, no-code interface with minimal setup costs and no ongoing maintenance requirements, it shows several important limitations when deployed in this domain. One of the most critical issues is its lack of transparency: the retrieval mechanism is essentially a black box, with no control over the ranking strategy or the number of documents retrieved. Moreover, it does not expose metadata or allow debugging, which prevents any form of traceability or post hoc validation of the results. This is especially problematic in regulated environments, where audits and documented evidence are essential. From a technical standpoint, the off-the-shelf solution is not designed to handle the type of semi-structured, highly technical documents typically found in pharmaceutical manufacturing, such as shift handover reports. It struggles with incomplete sentences, domain-specific abbreviations, and technical jargon. Furthermore, it does not support keyword-based retrieval, which is critical for handling structured identifiers such as batch numbers, often treated as out-of-vocabulary terms by generic models. It also fails to recognize date formats reliably, and it lacks mechanisms for aggregating information across documents, an essential capability in temporal or cross-shift analyses.

Nevertheless, the commercial solution has the merit of being lightweight in terms of resource allocation. It requires no domain adaptation, no data engineering, and no dedicated staff to maintain or supervise it, characteristics that might make it attractive for less critical or resource-constrained applications.

In summary, while the custom solution aligns with multiple principles of Human-

Centric AI, such as reliability, partial traceability, and human oversight, it falls short in terms of full transparency and explainability. Conversely, the off-the-shelf solution is easier to adopt and maintain, but fundamentally misaligned with the needs of complex, high-stakes industrial environments.

The Table 6.2 summarizes the comparison of the two solutions:

Dimension	**Custom Solution**	**Off-the-Shelf Solution**
Performance (Relevance, Faithfulness, Coverage)	High accuracy, reliable, consistent output	Lower accuracy, inconsistent results, many low scores
Explainability & Transparency	Partially traceable but not fully explainable; requires technical understanding	Black-box model; no transparency or traceability
Control and Customization	Fully controllable (retrieval strategy, document scope, parameters)	No control over retrieval logic or document processing
Technical Adaptability	Designed for structured, technical industrial documents	Not optimized for semi-structured or technical language
Handling of Domain Features (e.g. typos, codes, dates)	Supports keyword matching, handles domain codes	No keyword search; struggles with batch codes, dates, and technical formats
Aggregation Capabilities	Can summarize and synthesize across multiple documents	No aggregation or temporal reasoning
Implementation Complexity	Requires infrastructure (cloud/on-premise), setup effort, and maintenance	Ready-to-use, low technical entry barrier
Cost and Resources	High time, cost, and personnel effort	Low cost, no-code, no human supervision needed
User Interface	Requires front-end development and system integration	Includes usable front-end, immediately accessible
Trust and Adoption	Trust built through superior performance, but requires technical endorsement	Easier initial adoption due to simplicity, despite lower quality

Table 6.2: Comparative analysis between the Custom and Off-the-Shelf solutions across key dimensions.

In summary, while the custom solution aligns with multiple principles of Human-Centric AI, such as reliability, partial traceability, and human oversight, it falls short in terms of full transparency and explainability. Conversely, the off-the-shelf solution is easier to adopt and maintain, but fundamentally misaligned with the needs of complex, high-stakes industrial environments.

6.6 Further Developments

Given all the limitations and challenges widely discussed in section 6.5, this project represents an attempt to push innovation into a complex and traditionally rigid domain, where technological adaptation is often held back by strict constraints and regulatory frameworks. However it presents several lacking characteristics which prevent it to be fully explainable, easy and ready to be implemented, and scalable in a smooth way.

While the current system partially supports traceability of retrieved sources, future work should focus on making the inner working of the algorithm in the decision of which documents to retrieve more interpretable and accessible to all users. *Explainable AI is crucial for an organization in building trust and confidence when putting AI models into production* (IBM n.d.). One potential direction is the integration of explainability features such as highlighting which parts of the input documents contributed to the final output or displaying confidence scores associated with each document retrieved (Doshi-Velez and Kim 2017). Visual summaries or saliency-based explanations could assist users in understanding why specific documents were selected and how information was synthesized (Kares et al. 2025).

A second area of development refers to the evaluation process which can be potentially extended. The current assessment has demonstrated the higher performance of the developed system with respect to an off-the-shelf solution, however the set of test queries could be increased and varied even more, in order to evaluate the system in the most organic way as possible. Moreover, the system could evolve to support a more interactive and ongoing evaluation process. Expert users could share feedback directly through the interface, making it easier to point out missing information or

suggest improvements. This feedback would become part of the agent's workflow, helping the system learn and improve over time. In this way, evaluation wouldn't be just a one-time task, but a built-in quality check that continuously helps the system stay aligned with users' needs and expectations.

Another relevant direction for future development emerged directly during the implementation phase, through continuous engagement with stakeholders and end users. A recurring insight highlighted the need to standardize the format of shift handover reports across the different production lines. Although the current system is designed to handle free-form, semi-structured documents, introducing a more consistent and structured design would significantly enhance both its interpretative capacity and its output quality. One concrete proposal involves the inclusion of a predefined list of machines within the report template, allowing the person compiling the report to fill in dedicated fields for each machine. These fields would remain optional and flexible, enabling operators to describe events in free text, while still giving the system clear semantic signals. If a field contains information, the system understands that something noteworthy occurred for that machine; if left blank, it can safely infer the absence of relevant events. This change would improve the clarity and completeness of information encoding, facilitating more accurate retrieval, aggregation, and summarization. Importantly, this improvement also supports a broader strategy of stakeholder engagement as a driver of system robustness. Actively involving operators, supervisors, and engineers in the co-design of knowledge capture tools ensures that the AI system evolves in line with real operational needs. In this sense, standardization does not limit flexibility, but rather enhances efficiency and interpretability, ultimately leading to a more trustworthy and effective knowledge management solution.

So far, the system has been applied to the packaging and device module of the production line. However, its structure is flexible enough to be expanded to other parts of the manufacturing process, like upstream operations, formulation, or quality control. Even areas beyond production, such as laboratory work, logistics, or preventive maintenance, could benefit from a similar setup. Thanks to its modular design, this agentic architecture has the potential to support knowledge management

across many different areas where capturing and using information effectively really matters.

Another valuable direction for development is helping the system spot trends in operations and highlight when something unusual happens. By looking at past shift reports, it could learn to recognize patterns, like recurring alarms or long downtime, and automatically flag them to the right people. This feature would become even more powerful if the system were connected with the digital tools the company already uses, such as production dashboards, or internal notification tools. With this kind of integration, the system would go beyond simply retrieving information—it would become an active support in real-time decision-making.

In conclusion, to make the system easier to use and more accessible day-to-day, one key next step would be building a dedicated web application. A centralized and intuitive interface would allow users to submit questions, review the documents the system retrieves, read the generated summaries, and give feedback, all in one place. This kind of front end would not only improve the user experience, it would also make it possible to add important features like user access controls, activity tracking, and monitoring tools. Without a user-friendly interface, all the effort invested in developing the underlying architecture would fail to deliver tangible value to end users. These additions would help align the system with enterprise requirements and support its long-term use and scalability within the organization.

Chapter 7

Conclusion

Pharmaceutical industry is characterized by a considerable amount of regulation and complexity in operations. In this context knowledge management is not only a support function, but also a core component of efficiency, continuous improvement and compliance. Nevertheless, it remains a persistent challenge effectively capturing, structuring and making usable the knowledge generated on a daily basis on the production floor. Informal communication channels, such as emails, legacy systems, and fragmented documentation practices often contribute to a significant loss of operational knowledge. This issue is especially visible in shift handover processes, where valuable information is captured in unstructured or semi-structured documents that remain in employee's email inboxes and their retrieval is inefficient and time-consuming. While these reports are typically read by the following shift, the information they contain is rarely reused beyond the immediate operational context.

This thesis demonstrated the feasibility, and the complexity, of designing and deploying an AI-based system to support operational KM in such a context. The architecture developed can be classified as a Retrieval-Augmented Generation (RAG) system, designed to retrieve relevant production shift handover documents and generate structured summaries. What distinguishes this implementation is its agentic orchestration, built with the LangGraph library. By structuring the workflow as a graph of interacting nodes (retrieval, user selection, summarization), the system enhances modularity, traceability, adaptability, and the possibility of integrating different pipelines depending on the query of the user. These are essential properties

in environments where human oversight and compliance are central.

Yet, perhaps an important insight from this project is that technology alone is not sufficient. Successfully integrating a system of this kind requires more than technical deployment; it demands a process of organizational engagement and cultural transformation keeping. As Gupta, Iyer, and Aronson (2000) highlight, one of the most critical barriers to effective knowledge management is not the lack of tools, but the difficulty in securing stakeholder involvement in changing their processes. Involving end-users early, through participatory design sessions, continuous feedback, and iterative prototyping, is therefore essential.

In fact, during the development phase of this project, people coming from the shop floor, potential users of the system, understood the potential value of this system and voluntarily began to propose a modified version of shift handover reports, adding more structured and machine-specific annotations. This spontaneous behavioral shift, triggered by exposure to the prototype, suggests that co-design and stakeholder inclusion can catalyze meaningful cultural change, which in turn increases the future utility and scalability of AI-powered tools.

Another key takeaway relates to the regulatory and operational constraints of introducing AI in manufacturing. Any AI system must be transparent, controllable, and auditable, characteristics that are difficult to guarantee with black-box models. This leads to an inherent trade-off between:

- Custom-built solutions, which are more expensive and complex to develop, but offer higher levels of control, contextual fit, and potential transparency;
- Off-the-shelf general-purpose models, which are easier to deploy and scale but often underperform in domain-specific tasks and may not align with compliance requirements.

Choosing the right approach depends on the organization's priorities, technical capacity, and risk tolerance. However, this thesis advocates that in high-stakes industrial domains, a tailored approach, despite its cost, can unlock deeper integration, long-term value, and trust from end-users.

Ultimately, the broader reflection emerging from this work is provocative: while industries are increasingly investing in AI and automation technologies, the effectiveness of such tools is constrained by outdated information ecosystems and organizational paradigms. For KM to truly evolve, a systemic shift is needed: one that goes beyond the deployment of new tools and embraces new ways of working, sharing, and learning.

In conclusion, this thesis contributed both a concrete technological architecture and a set of organizational insights. It showed that AI-enabled KM is possible, but only when the solution is context-aware, user-centered, and culturally embedded. Future work could expand the system's capabilities and explore its integration with digital platforms or enterprise knowledge graphs. Yet the most important challenge ahead is not technical: it is the challenge of designing sociotechnical systems that bridge human expertise and machine intelligence, seamlessly, responsibly, and meaningfully.

Bibliography

Amjad, M. (Oct. 2024). *Advanced Querying Techniques with ChromaDB and Python: Beyond Simple Retrieval*. https://medium.com/@mehmood9501/advanced-querying-techniques-with-chromadb-and-python-beyond-simple-retrieval-c189a228c0a3. Accessed: 2025-04-07.

AutoGPT (Sept. 2025). *AutoGPT Platform*. https://agpt.co/. Accessed: 2025-09-21.

Brown, Tom and et al. (2020). "Language Models are Few-Shot Learners". In: *arXiv preprint arXiv:2005.14165*.

Cambridge University Press (n.d.). *Knowledge*. https://dictionary.cambridge.org/dictionary/english/knowledge. Accessed May 15, 2025.

Chroma (2024). *Chroma: The AI-native open-source embedding database*. https://www.trychroma.com/. Accessed: 2025-04-08.

CrewAI (Sept. 2025). *CrewAI Platform*. https://www.crewai.com/. Accessed: 2025-09-21.

Davenport, Thomas H. (2025). "Managing the Unstructured Data Challenge". In: *Harvard Business Review*. Online Article.

Desai, Akash A. (Dec. 2023). *Hybrid Search: Combining BM25 and Semantic Search for Better Results with Langchain*. https://medium.com/etoai/hybrid-search-combining-bm25-and-semantic-search-for-better-results-with-lan-1358038fe7e6. Accessed: 2025-04-08.

Doshi-Velez, Finale and Been Kim (2017). "Towards A Rigorous Science of Interpretable Machine Learning". In: *arXiv preprint arXiv:1702.08608*. URL: https://arxiv.org/abs/1702.08608.

Eschbach (2023). *Frequently Asked Questions on Pharma-Specific Knowledge Management*. https://eschbach.com.

Foundation, Python Software (2024). *re — Regular expression operations*. https://docs.python.org/3/library/re.html. Accessed: 2025-04-08.

Gupta, B., L. S. Iyer, and J. E. Aronson (2000). *Knowledge Management: Practices and Challenges*.

Harchaoui, Mohamed El (2023). *RAG Evaluation Metrics Explained: A Complete Guide*. https://medium.com/@med.el.harchaoui/rag-evaluation-metrics-explained-a-complete-guide-dbd7a3b571a8. Accessed: 2025-05-27.

Harsoor, Sharan (Nov. 2024). *Embeddings: A Deep Dive from Basics to Advanced Concepts*. https://medium.com/@sharanharsoor/embeddings-a-deep-dive-from-basics-to-advanced-concepts-f092765476fc. Accessed: 2025-04-08.

Hayes, Robert H. and Steven C. Wheelwright (1984). *Restoring Our Competitive Edge*. New York, NY: Collier Macmillan.

Hill, Terry J. (1987). "Teaching manufacturing strategy". In: *International Journal of Operations & Production Management* 6, pp. 10–20.

IBM (2024). *LangGraph Documentation*. https://langgraph.ibm.com/docs.

— (n.d.). *Explainable AI*. Accessed: 2025-06-04. URL: https://www.ibm.com/think/topics/explainable-ai.

IDC (2018). *The High Cost of Not Finding Information*. Tech. rep.

Kares, Felix et al. (2025). "What Makes for a Good Saliency Map? Comparing Strategies for Evaluating Saliency Maps in Explainable AI (XAI)". In: *arXiv preprint arXiv:2504.17023*. URL: https://arxiv.org/abs/2504.17023.

LangChain (2024). *LangChain: Building Applications with LLMs Through Composable Components*. https://github.com/langchain-ai/langchain. Accessed: 2025-04-08.

LangChain AI (2024). *LangGraph*. https://langchain-ai.github.io/langgraph/. Accessed: 2025-04-09.

Lewis, Patrick and et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *NeurIPS*.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press.

OpenAI (2022a). *Embeddings Documentation*. https://platform.openai.com/docs/guides/embeddings.

— (2022b). *text-embedding-ada-002*. https://platform.openai.com/docs/guides/embeddings. Accessed: 2025-04-08.

— (2023). *GPT-4 Technical Report*. https://openai.com/research/gpt-4.

— (2024). *GPT-4o: OpenAI's New Multimodal Model*. https://openai.com/index/gpt-4o. Accessed: 2025-04-08.

— (2025). *DALL·E Image Generation Tool*. https://openai.com/dall-e. Images generated by the author using DALL·E. Accessed: 2025-04-04.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *EMNLP*.

Ribeiro, V. B. et al. (2022). "Knowledge management and Industry 4.0: A critical analysis and future agenda". In: *Gestão & Produção* 29, e5222. DOI: 10.1590/1806-9649-2022v29e5222. URL: https://doi.org/10.1590/1806-9649-2022v29e5222.

Robertson, Stephen and Hugo Zaragoza (2009). "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Foundations and Trends in Information Retrieval* 3, pp. 333–389.

Schmager, Stefan, Ilias O. Pappas, and Polyxeni Vassilakopoulou (2025). "Understanding Human-Centred AI: A Review of Its Defining Elements and a Research Agenda". In: *Behaviour & Information Technology*. Published online: 16 Feb 2025. DOI: 10.1080/0144929X.2024.2448719.

Shaw, Duncan and John Edwards (2006). "Manufacturing knowledge management strategy". In: *International Journal of Production Research* 44.10, pp. 1905–1922. DOI: 10.1080/00207540500431339.

Symestic GmbH (n.d.). *What is Knowledge Management?* Accessed May 15, 2025. URL: https://www.symestic.com/en-us/what-is/knowledge-management.

Vector Stores | LangChain (2024). https://python.langchain.com/docs/concepts/vectorstores/. Accessed: 2024-04-08.

Wei, Jason and et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *arXiv preprint arXiv:2201.11903*.

Xu, Tony Li (2023). *Prompt Engineering: Role Prompting*. https://tonylixu.medium.com/prompt-engineering-role-prompting-7f757180011b. Accessed: 2025-06-05.

Yenduri, Gokul et al. (2023). "GPT (Generative Pre-trained Transformer) – A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions". In: *arXiv preprint arXiv:2305.10435*. DOI: 10.48550/arXiv.2305.10435. URL: https://arxiv.org/abs/2305.10435.