

d-fine



AI based Knowledge Management in Industry

Efficient information retrieval with
Large Language Models

Content

AI based Knowledge Management in Industry,
November 2023
© d-fine GmbH

1. Introduction	Page 3
2. Corporate knowledge management – a constant challenge	Page 3
3. LLMs in industry	Page 4
4. Example: Searching technical literature	Page 6
5. What is your use case – and how can we help?	Page 7
6. Getting off the ground with generative AI	Page 8

01.

Introduction

In an ever-more digitalized economy corporates from all sectors create and consume an ever-growing amount of data during their daily operations as the means to acquire, store and (potentially) analyze them become a commodity. It is common sense that leveraging these data to gain deeper insights and drive informed decision-making is a key factor in achieving a competitive advantage. Still, it requires both discipline in managing raw data and robust solutions and workflows to translate them into useful knowledge. This white paper explores how organizations can utilize advanced Large Language Models (LLMs), such as ChatGPT, to extract valuable insights from their Corporate Body of Knowledge (CBK) and leverage their frequently unexploited know-how pool to overcome resource shortage and gain efficiency in their operations.

02.

Corporate knowledge management – a constant challenge

Corporate know-how varies widely in content, form and in perceived criticality across industrial sectors but also within each company. However, the technical methodologies employed for information storage are seldom selected with future accessibility in mind, often resulting in accumulations of document piles and disjoint data silos that are rarely utilized actively.

“Imagine a situation where technical reports, which could simplify everyone’s life in the maintenance team, were centrally accessible in any situation and in almost any language. This is a scenario where the corporate body of knowledge actively benefits the corporate value chain.”

– Frederick Blumenthal, Lead Expert Generative AI

Beyond intellectual property in the proper sense, the CBK comprises i.a:

- Asset data (handbooks, maintenance documentation)
- Product and production data (recepties, line configurations, best-practices, schedules, load information, deadlines...)
- Governance data (back-office & IT process documentation, security & compliance guidelines, training material and working instructions)
- Supplier data (contractual information, availability, quality issues)
- Customer data (meta-data, complaints, feature requests)¹

In discussions with clients from finance, energy, healthcare, manufacturing and other sectors we have identified a set of universal issues that arise from such a diverse know-how basis:

- **Diversity:** Know-how is kept in multiple systems and in different format and languages and cannot be contextualized easily across these sources

¹ We intentionally do NOT consider employee data or personal customer data here. For an infrastructure concept that is particularly tailored to analyses of highly sensitive data see e.g., <https://www.eurodat.org/>

- **Quality:** Outdated information is not removed or updated – frequently because there are either no processes or resources
- **Accessibility:** Interested parties frequently just don't know that relevant information exists – and where to find it!
- **Verifiability (associated with data quality):** The degree to which one can rely upon “retrieved” information varies

03.

LLMs in industry

3.1

Selected use cases

ChatGPT and other LLMs are NOT going to magic all these issues away – in particular, they will not compensate data quality problems and sloppy data management processes. However, if properly set up, they simplify a few important tasks:

- Consolidation and contextualization of information provided in different locations, formats, and languages
- Centralization of access to data without the need to know their precise location
- Information retrieval in natural language with a substantially lowered “barrier to ask”, faster results and better user acceptance
- Substantiation of search results by reference to the original data resulting in improved trust into the AI

From our project experience we have derived a set of use cases for LLMs that we deem beneficial and realistic across all sectors but in particular for industry clients. In Figure 1 we locate these use cases on an (ordinal) matrix indicating our assessment of their Feasibility and their Added Value – both of which may vary on a case-by-case basis and are especially driven by the availability of data²

3.2

Typical Challenges

Neither are LLMs a cure for all data management topics, nor do they “come for free”. Hurdles are technical in nature or relate to governance and acceptance of the “digital colleague”.

Technical challenges are similar to those of other IT systems:

- To facilitate training of LLMs and searching their contents afterwards data sources must be interfaced – obviously dependencies, effort and complexity increase if multiple sources are required.
- Maintaining the system up to date requires a mature ML-Ops process spanning the entire model life cycle from training and testing, to deployment, maintenance, usage and decommissioning.
- Most importantly – and frequently also most painfully – to keep the system useful over time the quality of the data entering the training and usage phase must be ensured.

² Note that generative usage scenarios like coding or technical design assistants have technical (and security) intricacies that are beyond the scope of this whitepaper.

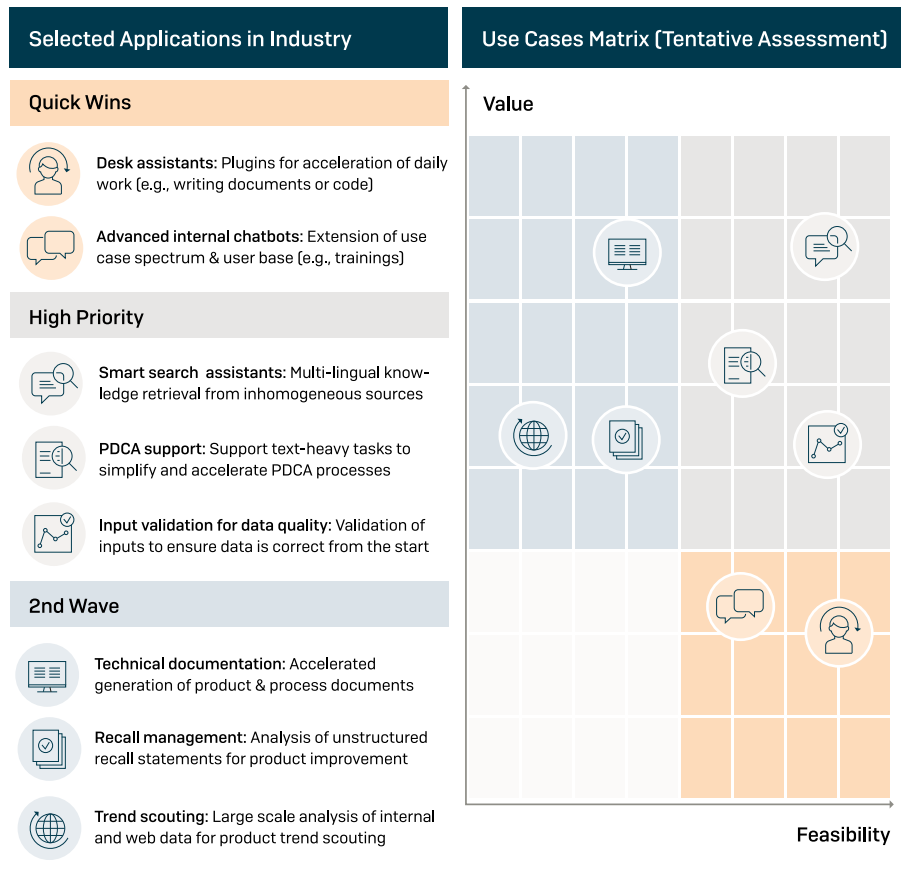


Figure 1: Assessment matrix for various industry related LLM use cases

On the governance side three things prove important:

- a clear policy what reference architecture to use and how to host an LLM within on-prem or cloud-borne corporate IT structures
- a very conscious decision which data to use and which ones NOT for training and retrieval in line with data protection regulations
- a concept how to allocate and manage permissions for searching for a given piece of information

In terms of user buy-in a widespread acceptance requires:

- user training and a clear communication of the possibilities and limits of the new assistance systems
- policies for their usage
- means to verify the proposed answer of the smart search assistant

To provide our clients with a test bed and to give a hands-on impression we have set up a simplistic demonstrator system capable to search corporate information items in almost any spoken language. Specifically it answers questions related to a set of user defined documents that are stored on a Microsoft Azure Search-Index beforehand. The system does not require training or hosting ones own LLM and is based on ChatGPT 4. Figure 2 shows an example, where we have queried the Open Access book “Elements of Robotics³” by Ben-Ari and Mondana which is available under creative commons license⁴ for an explanation of odometry - a common navigation method for low-cost robots.

Despite its out-of-the box character after a little prompt engineering the system delivers correct answers, and in most cases also correct references to their origin in the searched documents. The latter feature is important insofar, as a critical user MUST challenge the results presented to him and should be careful to trust the AI’s answers indiscriminately.

“LLMs help to make the corporate body of knowledge accessible and useable in a practical way, e.g., via generation of worker guidance for maintenance tasks on the shop floor.”

– Tassilo Christ, Partner d-fine Industrial Solutions

³ Ben-Ari,M.and Mondala, F.: Elements of Robotics, Springer Open, 2018; <https://doi.org/10.1007/978-3-319-62533-1>

⁴ <http://creativecommons.org/licenses/by/4.0/>

The screenshot displays the 'Smart Search Demo App' interface. At the top, there is a navigation bar with 'Smart Search Demo App', 'Chat', 'd-fine', and 'Smart Enterprise Data Management'. Below the navigation bar, there is a search prompt: 'Briefly, in several sentences describe the context of odometry for robot navigation'. The search results are displayed in a list of citations, with two entries highlighted. The first entry is selected, and its content is shown in a detailed view on the right. The detailed view shows the title '5.4 Navigation by Odometry' and the beginning of the text: 'Suppose that you are in a car and your navigation system issues the following instruction: "Go 700m turn right." Now your task is very simple. Make observations of your car's odometer which measures how far you have traveled. When its value approaches 700m beyond its initial reading, look for a street on the right. An odometer in a car measures speed and time, and multiplies the two values to compute the distance traveled.' The interface also includes a search bar at the bottom left with the text 'Type a new question' and a 'gpt4' dropdown menu.

Figure 2: Result of searching a textbook for explanation of a navigation concept. The system presents a summary including information from multiple text sections.

What is your use case – and how can we help?

Off-the-shelf AI solutions that tackle “the typical” use cases and need “only a little” configuration rarely provide a long term competitive edge. According to our experience harvesting sustainable benefits instead requires a bespoke and focused approach that combines deep domain expertise from the client side and technological and methodological insights injected by a team of specialists. This is even more the case with AI topics as the underlying technology requires a thorough mathematical background to deliver measurable added value.

d-fine has an extensive track record in co-creating customized LLM-solution with clients from finance, manufacturing, and other industries. Going way beyond the ubiquitous “AI-PoCs” we support our customers to bring solutions into production. Figure 3 shows a selection of successful AI projects in which we have delivered NLP-systems that are tailored to specific customers needs.

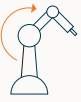





<p>Manufacturing Industry</p>  <p>Shop floor assistant: Smart search tool-stack based on MS Azure OpenAI services to simplify information retrieval from technical documents</p>	<p>Medical Research</p>  <p>Information extraction: Automated text-based research on device safety to support the expert work for a medical device manufacturer</p>
<p>Cloud Provider</p>  <p>Customer chatbot: Data pipelines for enrichment of ChatGPT with internal documents; Backend / API development and prompt engineering</p>	<p>IT Operations</p>  <p>Document quality screening: Automated analysis and monitoring of internal IT documentation for completeness and quality</p>
<p>Public Sector</p>  <p>Scientific text analysis: Information retrieval research project with the European Food Safety Authority; Integration of GPT-3-API and prompt engineering</p>	<p>Banking Regulation</p>  <p>AI-augmented search: LLM-supported search platform for information retrieval from a large document corpus on banking regulation</p>

Figure 3: Result of searching a textbook for explanation of a navigation concept. The system presents a summary including information from multiple text sections.

To gain an understanding of the potential and the limitations of generative AI – and to avoid disappointment – one must identify and evaluate business cases thoroughly. Frequently our projects therefore start with a dedicated use case ideation phase or workshop to which we contribute our broad cross-sectoral experience. We help our clients to identify applications for LLMs and to assess technical feasibility, implementation cost and potential impact. To verify the value of generative AI and to facilitate a steep learning curve we recommend to implement a selected high-value pilot use case and to extend the solution to other processes and in an agile manner subsequently.

Our offer to you: We are very excited to identify which use cases apply to your company! Let us find out how ChatGPT & Co. can deliver added value to you in a workshop in which we demonstrate how our smart search tool-stack performs with your data!

Authors



Frederick Blumenthal
Manager and Expert NLP Solutions
d-fine GmbH, Munich
Frederick.Blumenthal@d-fine.com



Dr Ulf Menzler
Senior Manager and Expert MLOps
d-fine GmbH, Dusseldorf
Ulf.Menzler@d-fine.com



Dr Fedor Petrov
Manager and Expert Manufacturing Systems
d-fine GmbH, Frankfurt
Fedor.Petrov@d-fine.com



Dr Tassilo Christ
Partner and Head of Industrial Solutions
d-fine GmbH, Munich
Tassilo.Christ@d-fine.com

Berlin

d-fine GmbH
Kranzler Eck
Kurfürstendamm 21
10719 Berlin
Germany
berlin@d-fine.de

Dusseldorf

d-fine GmbH
Dreischeibenhaus 1
40211 Dusseldorf
Germany
duesseldorf@d-fine.de

Frankfurt

d-fine GmbH
An der Hauptwache 7
60313 Frankfurt
Germany
frankfurt@d-fine.de

Hamburg

d-fine GmbH
Am Sandtorpark 6
20457 Hamburg
Germany
hamburg@d-fine.de

London

d-fine Ltd
14 Aldermanbury Square
London, EC2V 7HR
United Kingdom
london@d-fine.co.uk

Milan

d-fine s.r.l.
Via Giuseppe Mengoni 4
20121 Milano MI
Italy
milano@d-fine.com

Munich

d-fine GmbH
Bavariafilmplatz 8
82031 Grünwald
Germany
muenchen@d-fine.de

Stockholm

d-fine AB
Nybrogatan 17
114 39 Stockholm
Sweden
stockholm@d-fine.se

Utrecht

d-fine BV
Stadsplateau 7
3521 AZ Utrecht
Netherlands
utrecht@d-fine.nl

Vienna

d-fine Austria GmbH
Seilerstätte 13
1010 Vienna
Austria
wien@d-fine.at

Zurich

d-fine AG
Brandschenkestrasse 150
8002 Zurich
Switzerland
zuerich@d-fine.ch