Fraunhofer ISST

Fraunhofer Institute for Software and Systems Engineering ISST

Data Spaces and Foundation Models

Enabling High-Quality Artificial Intelligence

Management Summary

How can a combination of data spaces and foundation models enable high-quality artificial intelligence — and who benefits from it?

Artificial intelligence (AI) is becoming a key driver of competitiveness across industries, enabling companies in manufacturing, finance, healthcare and other sectors to streamline operations, improve decision-making and create new business models. However, the effectiveness of AI depends on access to high-quality, industry-specific data — a challenge as most AI models, including foundation models, are primarily trained on public internet data. Regulatory requirements such as the EU AI Act further complicate the landscape, demanding transparency and compliance in AI-driven decision-making.

Despite their capabilities, foundation models face significant limitations due to their reliance on unstructured, publicly available data. Many lack industry-specific knowledge, struggle with data quality and bias, and create compliance risks due to unclear data sources. Current solutions, such as custom AI training, are often costly, time-consuming, and limited by a lack of access to real-world company data. As a result, businesses that rely solely on off-the-shelf AI models risk losing their competitive advantage, facing legal challenges, and suffering from operational inefficiencies due to inaccurate or outdated AI-generated insights.

To fully harness the potential of AI while ensuring data security, sovereignty, and compliance, companies need a trusted infrastructure for controlled data exchange. Data spaces provide a solution by enabling secure and standardized data sharing, allowing organizations to retain full control over their proprietary information while making it accessible for AI-driven innovation. By integrating data spaces with foundation models, businesses can unlock new opportunities, ensure regulatory compliance, and enhance AI performance. This white paper explores practical approaches such as retrieval-augmented generation (RAG) and fine-tuning, which allow AI models to incorporate trusted, industry-specific knowledge, making them more reliable and effective for real-world applications.

Contents

Authors

Dr. Marcel Altendeitering

Sebastian Becker

Maurice Boiting

René Brinkhege

Dr. Tobias Guggenberger

Maik Mannsfeld

Michael Steinert

Daniel Tebernum

1. The Data and AI Value Chain

Data is of critical value to industry and businesses, serving as a strategic organizational resource that can enable informed decision-making and data-driven innovation. Data can increase operational efficiency, unveil market trends and insights, and lead to better customer experience, to name just a few benefits. This is new to many traditional companies, which are confronted with a fundamental change as the delivery of data-driven services and products becomes increasingly important[1]. Companies that can effectively leverage data resources and have robust data management practices in place are better positioned to gain competitive advantages and ensure sustainable growth[2].

Artificial Intelligence (AI) has found its way into the corporate landscape with the introduction of ChatGPT in particular and is one of the megatrends of this decade[3]. AI describes all efforts to imitate a machine's ability to perform human skills, such as logical thinking, learning, planning and creativity, and can be divided into various sub-areas including machine learning, deep learning, natural language processing and knowledge representation[4]. AI offers great potential for realizing cost savings[5] and productivity improvements, thereby becoming a major source of competitive advantage[6].

All of these tasks require several key resources. First, AI needs high-quality, well-managed data, or as summarized by Gröger (2021): "There is no AI without data"[7]. Second, AI needs IT resources, such as computing powers and intelligent algorithms, which also requires the right talents. Third, AI needs funding and energy to promote the development of AI solutions.

Yet, Europe is currently not leading in the relevant domains, and AI resources (data, talent, etc.) are highly distributed. To leverage these benefits of AI and increase global competitiveness, European organizations must spur AI innovation and develop a vision for trustworthy and sustainable AI. In light of this, it is most important that organizations utilize their "data treasure," which holds enormous potential for industrial AI applications and foundation models. Data spaces offer one way to use the private data held by organizations. This white paper examines how integrating data spaces with foundation models can enhance AI efficiency.

2. Related Work

2.1. Introduction to AI and Foundation Models

The proliferation of AI is based on the success of foundation models (FMs). FMs are large neural deep learning networks that have changed the approach to AI. They are AI models designed to produce a wide and general variety of outputs and are capable of a range of possible generative tasks and applications, such as text, image, and audio generation. Because FMs have been trained on many datasets for an array of application scenarios, they can be adapted to specific applications. As a result, existing FMs are accessed and adjustments are made instead of developing AI from scratch, which reduces a significant amount of the AI development effort[8].

As a subcategory of FMs, large language models (LLMs) are based on large text bodies and are used to generate texts. They can be standalone systems or can be used as a "base" for many other applications. Prominent examples of LLMs are OpenAI's GPT[9], Google's Gemini[10], Meta's LLaMA[11] series, and Teuken 7B[12] as a European alternative. Despite all the advantages of LLMs, there are currently inherent limitations to them, particularly when working with specific, up-to-date, or specialized information, but efforts are being made to overcome these limitations.

Specifically, two key techniques to enhance FMs and address these limitations are **retrieval-augmented generation (RAG)** and **fine-tuning**. Each of these methods offers distinct advantages and use cases, allowing FMs to be tailored to various applications while improving accuracy, relevance, and efficiency[13].

FMs are trained on a vast amount of data. The relevant units are tokens, into which the training data is converted. For text data, tokens are individual elements such as words, punctuation marks, or subwords. Tokenizers from established natural language processing (NLP) tools support this task[14]. GPT-3, for example, was trained on an estimated 45TB[15] of text data or approximately 499 billion tokens. The figures for GPT-4 are not publicly known but are estimated to be much higher. Meta's largest model, LLaMA 3.1 405B, which has capabilities comparable to GPT-4, was trained on 15 trillion tokens. Nowadays, not only is text data tokenized, but image and sound data as well, leading to what are called "multimodal models".

All of these models are based on the transformer architecture[16]. A transformer works like this: First, the input tokens (for example, the question of the user plus context) are converted into numerical vectors called embeddings, which encode semantic and contextual information of the input. The transformer then tries to predict the next token based on the input data. A key component here is the attention mechanism, which evaluates which parts of the input data are relevant for the prediction of the next token. This attention is learned in the training process. Finally, the transformer predicts which token is the most likely to follow that specific sequence of tokens.

Researchers found that if models are trained with more tokens and have more parameters, they become much better at predicting the next token. In terms of making better predictions, there is an emergent ability to capture broader concepts that surpass mere modeling of just words or grammar. At this stage, the models are very good at predicting the next token, but they are not always as usable. For example, if you give them a question, it might happen that the most likely continuation of this sequence is more questions. For this reason, most models additionally go through a step called reinforcement learning from human feedback (RLHF)[17]. In RLHF, the outputs of the model are evaluated by humans, and the feedback is used to fine-tune the models, so they give responses that more closely align with human preferences. FMs are typically saved in a serialized format (e.g., HDF5, GGML) that allows them to be efficiently loaded and used for inference or further training.

Gathering the large amounts of data necessary for creating FMs can be problematic. Currently, commercial models are largely based on data gathered from the internet using web crawlers[18]. This approach comes with several problems, such as a lack of data quality and unclear legal frameworks[19]. The following table summarizes the challenges that FMs currently face.

Challenges	Description
Unclear legal framework	Companies face problems in gathering huge amounts of training data in a legally compliant way. Current models use publicly available data (e.g., texts, images) as a basis for training. The EU AI Act, which comes into effect in Q2 2026, mandates transparency over AI systems and the data used for training.
Emerging "data winter"	The MIT Data Provenance Initiative found that data availability is drying up due to data providers blocking web crawlers or setting up paywalls. As a result, around 5% of the data used in C4 (Colossal Clean Crawled Corpus) is no longer available.
Unreliability and bias	Bias is a distinct possibility as models can pick up false information, hate speech and inappropriate undertones from training datasets.
Industry-specific data	FMs neglect industry-specific data such as machine data or high-quality texts. Leveraging industry-specific data for the training of FMs can help provide reliable answers.
Data quality	The quality of the training data may vary, which can affect the reliability of the results.
Adaptability	Adapting models to specific applications or domains can be time-consuming and complex. LLMs have a fixed knowledge cutoff, limited to the data they were trained on up to a specific point. This makes it challenging for them to handle queries related to later events or insights discovered after their training period. LLMs are designed to be versatile, but this generality means they may not perform optimally in highly specialized domains without further adjustment.
High computational costs	Training large models requires significant computational resources and energy.
Hallucinations	LLMs might generate inaccurate or fabricated information, often referred to as hallucinations as they focus on ongoing conversations rather than correct information.

Table 1. Summary of Challenges of Foundation Models

2.2. Applications of AI and Foundation Models in the Industrial Context

With their ability to process large volumes of data and learn from complex patterns, AI, particularly foundation models (FMs), is driving significant innovation and optimization across various industries. In sectors like manufacturing, logistics, and maintenance, AI and FMs present numerous opportunities for process improvement and the creation of new business models. In industrial practice, various types of AI applications can be identified.

The following outlines five key types of AI applications:

1. Use of AI-based digital services
Industrial enterprises consume digital or smart services offered by external software vendors that leverage AI. Examples include AI-enhanced sensor services for monitoring equipment and estimated time of arrival (ETA) predictions in logistics.

2. Use of conventional AI for own digital services
Companies use their own or customer data to develop digital services for their clients. Examples include predictive maintenance and condition monitoring, where data generated during product use enables tailored solutions.

3. Use of foundation models by industrial enterprises
FMs find applications in "white-collar" domains, such as creating reports, supporting decision-making, or automating routine administrative tasks. These tasks leverage the capabilities of LLMs to streamline processes and enhance efficiency.

4. Enrichment of foundation models by industrial enterprises
By integrating private, enterprise-specific data with generative AI using techniques like retrieval-augmented generation (RAG), companies can significantly enhance the relevance and performance of LLMs. This approach allows enterprises to derive greater value from their proprietary data.

5. Shared industrial foundation models/LLMs
Collaboration between multiple organizations to jointly develop and fine-tune foundation models (FMs) or large language models (LLMs) represents a promising, yet underutilized, approach. Sharing data to train a shared model can unlock the potential of data collaboration, but it requires trust, data sovereignty, and appropriate tools to ensure compliance and mutual benefits.

As highlighted, successful implementation of AI and FMs in industry depends on access to high-quality, comprehensive datasets. This is where data spaces become essential. They play a crucial role in facilitating the secure and efficient exchange of data across multiple stakeholders and sectors, enabling organizations to share and link data while maintaining control and minimizing security risks, as further discussed in the next chapter.

2.3. Data Spaces

Data spaces are a new concept aimed at facilitating the secure, trustworthy, and efficient exchange of data between different organizations[20]. Innovation in the form of new products and services is often driven by data shared between individuals and organizations. In this sense, data spaces represent the "motor" of data ecosystems, which act as a key enabler for the transformation of whole industries towards an integrated digital economy[21]. The concrete business value of a data space lies in the standardization of the data exchange infrastructure, including standardizing interfaces and information models.

CEN CENELEC defines a data space as an "interoperable framework, based on common governance principles, standards, practices and enabling services, that enables trusted data transactions between participants."[22]

Using the definition, there are multiple key features relevant to data spaces[23]:

- **Interoperability & standards:** Interoperability in data spaces refers to the ability of different systems, platforms, and applications to communicate and exchange data. It is crucial to ensure that data can be utilized across various technologies and organizational boundaries, significantly simplifying interorganizational collaboration. A key element for realizing interoperability is the IDSA Dataspace Protocol[24]. This protocol defines the required schemas and interactions for cataloging data as well as negotiating contracts and usage agreements within a data space.

- **Trust & data sovereignty:** Data sovereignty means that data owners retain full control over their data when sharing data with external data consumers. Data sovereignty is achieved through clear access rights, data protection policies, and the ability to encrypt data. Open-source implementations such as the Eclipse Data Space Components (EDC)[25] framework implement data sovereignty concepts using common access patterns and data protection standards. To further improve trust in data spaces, Gaia-X[26] offers a reference architecture for federated and trusted data sharing using components such as the Digital Clearing House.

- **Governance:** Governance in data spaces involves establishing clear rules, guidelines, and procedures for data exchange and collaboration between participating parties. This aspect includes defining roles and responsibilities, complying with legal and regulatory requirements (e.g., EU Data Act), and setting standards and establishing processes for secure data handling. A well-defined governance framework is essential to ensure seamless data sharing and interorganizational collaboration.

- **Flexibility:** Flexibility is at the heart of data spaces and aims to accommodate a wide variety of different data sources and types, including structured and unstructured data. This flexibility enables the adoption of data space technologies to different industries and use cases, whether in mobility, healthcare, or logistics. As a result, companies can integrate data from diverse sources to create comprehensive data products and implement use cases across domains.

The fact that data spaces enable trustworthy data sharing between different organizations makes them valuable for AI and FMs. The following section delves into the potential combination of the two concepts and describes the benefits and architectures for integration.

3. Data Spaces and Foundation Models

3.1. The Contribution of Data Spaces

Data spaces provide a controlled environment where data can be shared in a secure and standardized manner without creating centralized data silos. At the same time, data spaces ensure data sovereignty, which means that each participant remains in control of their data products. As a result, data spaces create trust in data exchanges and help facilitate data sharing between companies and organizations. **Consequently, data spaces can serve as a valuable component in the FM architecture by providing secure access to reliable, high-quality, and privately owned data sets.**

The benefits of data spaces for FMs can be summarized as follows:

- **Data sovereignty**: Data spaces ensure that data owners maintain control over their data while providing the necessary information to train or aid the FMs with additional context information.
- **Improved data quality**: Through standardized processes within data spaces, only verified and quality-assured data is used in the models, leading to more accurate results.
- **Access to private data sets**: Data spaces can provide access to privately owned data sources (e.g., corporate data), offering FMs a broader knowledge base and improving their generative capabilities (including context-specific data, for example).
- **Collaboration and interoperability**: Data spaces facilitate the secure and efficient data exchange between different organizations, fostering collaboration and innovation in FM development.
- **Regulatory compliance**: Data spaces provide the necessary infrastructure to ensure that the data exchange complies with strict privacy regulations and legal frameworks.

3.2. Integrating Data Spaces and Foundation Models

Over time, FMs and LLMs have significantly increased in their capabilities, yet they predominantly leverage general knowledge. Contextual knowledge embedded in companies remains largely unused. Data spaces can help leverage these data sets and support the enrichment of establishment models or the collaborative development of FMs and LLMs (see categories 4 and 5 in Section 2.2).

Traditional methods of interacting with LLMs involve direct prompting. Users send a query, and the LLM generates a response based on the information encoded in its weights and the context provided within the prompt. This approach, while straightforward, is limited by the extent of the context knowledge inherent in the prompt and the LLM's weights.

There are two methods for addressing these limitations: retrieval-augmented generation (RAG) and fine-tuning. Both methods are intended to enhance the reliability and accuracy of FMs without re-training the model itself. They both represent examples of category 4 in Section 2.2. These two methods are described further below.

3.2.1. Retrieval-Augmented Generation (RAG)

The retrieval-augmented generation (RAG) method can enrich the prompt with additional context information from internal company data and other sources. In the simplest form of RAG, a middle component vectorizes the user's prompt and compares it with vectorized data from the company's knowledge base. The most relevant matches are then added to the prompt, which is forwarded to the LLM. This augmented prompt allows the LLM to generate a response that is better informed by the specific context of the company's internal knowledge[27].

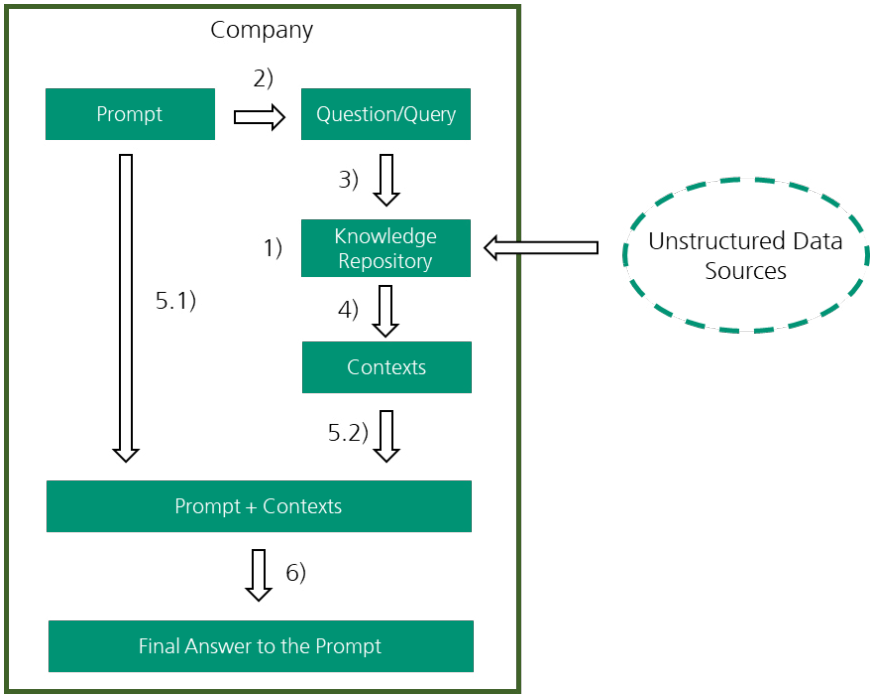

Figure 1: RAG workflow illustrates the process of augmenting prompts with external context.

The RAG process works in several stages:

- **Data provision and transformation**: First, external data that exists outside of the LLM's original training set is collected from a variety of sources such as documents (PDFs, DOCX, PPTX), wikis/intranets, corporate websites, knowledge graphs, and relational databases (e.g., SQL, PostgreSQL). It is converted into vector representations using embedding techniques such as Word2Vec[28] or BERT[29], These vectors are numerical representations of the data that are stored in the knowledge repository **1)** for future retrieval[30]. **User query**: The process begins when the user submits a prompt or question/query **2)**. This prompt is also transformed into a vector representation using the same embedding techniques (e.g., Word2Vec or BERT) to ensure that it can be effectively compared with the data stored in the knowledge repository.
- **Searching contexts**: Once the user's query is vectorized, a retrieval system (e.g., Pinecone[31] or FAISS[32]) compares the query's vector representation to the vectors stored in the knowledge repository **3)**.
- **Retrieving contexts**: The system identifies the most relevant contexts by calculating vector similarities, retrieving the pieces of data that are most closely aligned with the user's query[33] **4)**.
- **Contextualizing the prompt**: The retrieved contexts are combined with the original user prompt, thus augmenting the initial query with additional relevant information **5.1), 5.2)**. The prompt and contexts combination is then sent to the LLM.
- **Generating the final answer**: Finally, the LLM generates a response based on the augmented prompt, which is more accurate and contextually appropriate **6)**.

- Building on the concept of RAG, we can use an entire data space to enhance contextual knowledge. When a company sends a query, it is not only matched against its internal data but also against data available within the data space. Other companies can provide match scores for relevant knowledge, which the querying company can then decide to incorporate. If considered valuable, the additional knowledge is requested and used to further augment the prompt, leading to a more grounded response. The hypothesis is that the more context knowledge an LLM has access to, the higher the utility of its responses. This approach not only addresses the inherent limitations of current models but also opens new avenues for generating more accurate and contextually aware responses. The workflow of the data-space-augmented RAG is illustrated in Figure 2 below.

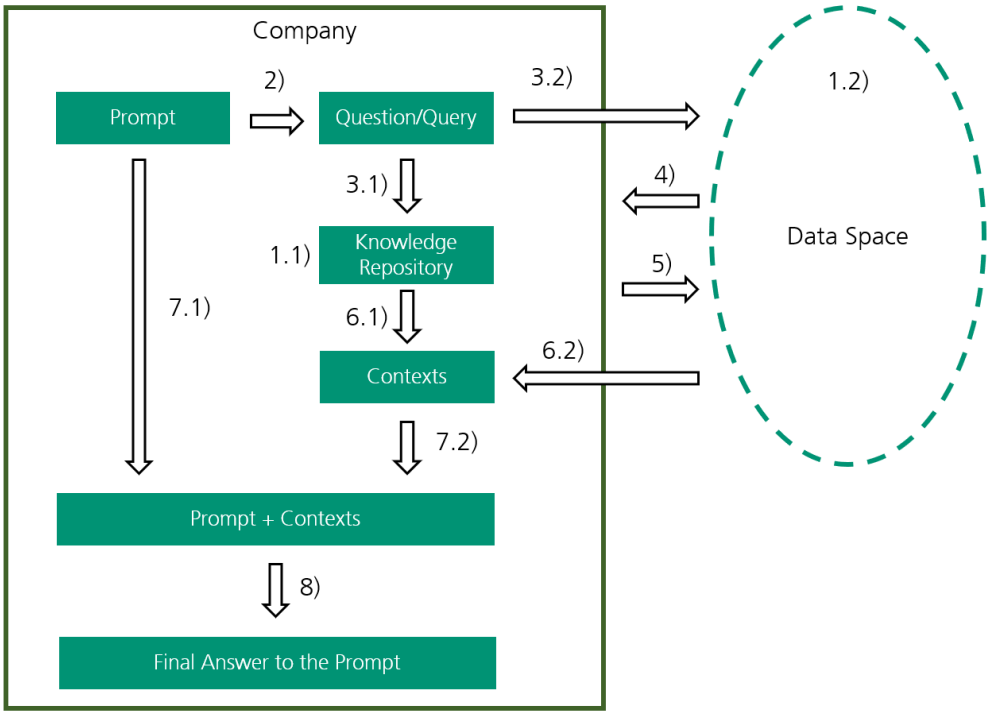

Figure 2. Data-space-augmented RAG workflow illustrates the process of augmenting prompts with context from a data space.

The data-space-augmented RAG process has some additional stages:

- **Data provision and transformation**: As with classical RAG, in the first step, external data that exists outside of the LLM's original training set is collected from a variety of sources. It is converted into vector representations. These representations are stored in the knowledge repository for future retrieval[34] **1.1)** This is not only done by one company but by multiple data space participants **1.2)**, each populating their own knowledge repository.
- **User query**: If the user now submits a prompt **2)**, the prompt is transformed into a vector representation.
- **Searching contexts**: Once the user's prompt is vectorized, a retrieval system compares the query's vector representation to the vectors stored in the knowledge repository **3.1)**. Additionally, the query is sent to the data space to find participants who also have relevant context information. They each take the query and compare it to their knowledge repositories **3.2)**.
- **Evaluate data space results:** If relevant context information is found in the data space, the metadata of the found context information is sent back to the requesting company. This metadata includes a score on how well the found context fits to the query **4)**.
- **Requesting data space results:** The requesting company can select and request the context data that it would like to use **5).**
- **Retrieving contexts**: The system identifies the most relevant contexts inside its internal

knowledge repository **6.1)** while also considering the requested context from the data space **6.2).**

- **Contextualizing the prompt**: From this step on, the process is the same as for the classical RAG: The retrieved contexts are combined with the original user prompt **7.1), 7.2)**.
- **Generating the final answer**: Finally, the LLM uses both its internal knowledge and the externally retrieved data (This time including context from the data space) to provide a final answer to the user's query, which is more accurate and contextually appropriate **8)**.

However, RAG has certain limitations. A key limitation is the LLM's context window, which defines the maximum amount of information the model can handle at any one time[35]. Even if RAG retrieves large amounts of relevant data, only some of it can be fed into the model if the context window is small. In addition, context management becomes critical in RAG, as overloading the context window with excessive information could degrade the quality of the response. This requires careful selection of the most relevant information to include in the query.

Another key consideration is keeping external data up to date. As data evolves, documents need to be updated asynchronously, and their embedding representations refreshed to ensure accurate retrieval. This can be achieved through automated real-time processes or regular batch updates. Addressing data freshness is a common challenge in data analytics, and various data science change management techniques can be effectively applied.

Despite these limitations, RAG offers significant advantages. It excels in scenarios where dynamic, up-to-date data is required, such as when an LLM is asked about recent developments in an industry or about events. Furthermore, by providing sources for the retrieved data, RAG increases the transparency of the model's output, which is valuable in applications where trust and verification are critical. Moreover, since RAG relies on external information to augment the LLM's responses, there is no need to retrain the model, which can be resource intensive.

RAG has a wide range of use cases, including customer support chatbots that need to be constantly updated with the latest product documentation, financial news services that provide real-time market updates and insights, and research assistance tools that can provide relevant academic papers or technical documents to support queries.

3.2.2. Fine-Tuning

Fine-tuning involves retraining an LLM on a smaller, specialized dataset to optimize it for a particular domain, task, or style. During this process, the model's weights are adjusted to incorporate new patterns, language, or tone specific to the dataset, making it more adept at handling queries in a particular context[36]. The fine-tuning process is illustrated in Figure 3 below.

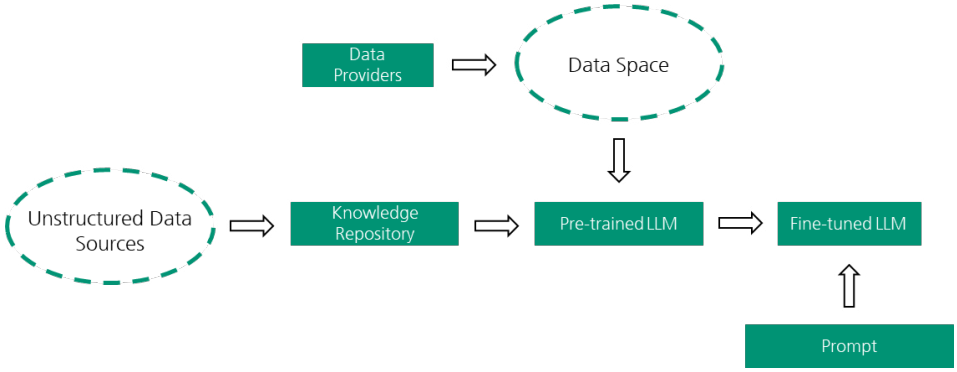

Figure 3. The fine-tuning process, where unstructured data from a data space is used to further train a pre-trained LLM, enabling it to generate more specialized responses.

The fine-tuning process begins with the selection of a domain-specific dataset tailored to a particular industry or task. This dataset could include legal documents, financial reports, or medical case studies. The data is first passed through a data space, where it is cataloged, negotiated and transferred from a domain-expert data provider or other data sources. Once this data has been collected, the pre-trained LLM uses it to undergo further fine-tuning, allowing the model to learn the specific nuances of the targeted domain[37]. After fine-tuning, the model, now referred to as a fine-tuned LLM, generates responses that match the vocabulary, tone, and structure of the domain.

Fine-tuning has several advantages. It allows the model to behave like an expert in a particular domain, providing specialized responses with the correct terminology and tone. This makes the model particularly valuable in industries such as law, where it can summarize contracts or provide legal insights. In addition, fine-tuning can improve the speed and efficiency of the model by optimizing it for specific tasks, allowing it to generate more concise and relevant responses even with smaller context windows. This reduces computational and inference costs[38]. Fine-tuning also enables the model to adopt specific communication styles, so that responses can be tailored to reflect an organization's brand voice for more personalized output.

There are many use cases for fine-tuning. For example, fine-tuned LLMs can be used as legal document summarizers, providing accurate, domain-specific summaries of complex legal texts. In customer service, fine-tuned models can adopt a company's tone and style to provide consistent responses, and in industry-specific research, fine-tuned LLMs can generate technical summaries or insights relevant to fields such as finance, healthcare, or education.

3.2.3. Combining RAG and Fine-Tuning

Both RAG and fine-tuning have advantages and disadvantages. The following table compares and summarizes the two concepts:

Features	RAG	Fine-Tuning
Data Requirements	Ideal for real-time data from dynamic sources	Works best with static or domain-specific data
Response Accuracy	Improves accuracy by pulling in external information	Highly accurate for specialized, predefined tasks
Context Limit	Limited by the model's context window	Not constrained by the context window once trained
Hallucination Risk	Reduced by providing relevant source information	Low, but limited to the dataset it was trained on
Adaptability	Quick to adapt to new information without retraining	Requires retraining to adapt to new data
Transparency	Can provide sources for the information used in responses	Responses are embedded into the model, without external citations
Cost & Speed	May have higher compute costs due to external data retrieval	More efficient during inference as the model is already fine-tuned

Table 2. Comparison of RAG and Fine-Tuning.

For many real-world applications, a hybrid approach combining RAG and fine-tuning offers a powerful solution that leverages the strengths of both methods. Fine-tuning allows a model to be highly specialized in a particular domain, ensuring that it adopts the correct language, terminology, and style. At the same time, RAG complements this specialization by enabling the model to access up-to-date external information, especially in rapidly changing environments.

Users of such combined systems have the advantage of leveraging three levels of knowledge: (1) the general knowledge provided by the foundation model, (2) the companies' inherent domain-specific knowledge inserted through fine-tuning, and (3) the latest and current knowledge inserted at response time through RAG. This dual approach is particularly effective in dynamic industries where accuracy and real-time data are critical.

One example: financial news service

Fine-tuning: The model is fine-tuned using financial terminology, historical market data and industry-specific reports. This enables the model to understand complex financial terms and generate summaries or insights with a high degree of accuracy. To achieve the best results, the model should gain access to private financial data using data space, which makes the inherent wisdom of an organization available to the AI model. The fine-tuned model can then be used by employees or customers and better utilize the company's knowledge.

RAG: At the same time, the model uses RAG to retrieve the latest financial news, stock prices and real-time market updates. This ensures that the model's responses are both relevant and timely, reflecting current market conditions and trends. A user who is prompting the model can thus make informed decisions and be sure that the model takes the latest developments into consideration.

This combination of RAG and fine-tuning creates a versatile and robust AI system that balances domain-specific expertise with the ability to incorporate real-time, dynamic information. Such a hybrid approach is ideal for building specialized AI applications in areas such as finance, healthcare, law, and customer service, where both accuracy and up-to-date information are critical to success.

3.3. The Contribution of Data Trustees

Data trustees play a key role in ensuring that sensitive and personal data can be accessed in a secure and ethical manner for the development of FMs. As neutral intermediaries, they provide an essential layer of trust and compliance, which is critical in an era of growing privacy concerns. By facilitating the controlled exchange of data, data trustees help to address some of the key challenges facing FMs today (see „3.1. The Contribution of Data Spaces" on page 9).

Data trustees provide secure and controlled access to data. They act as intermediaries, ensuring that data is only accessed and used under strictly regulated conditions[39], particularly when training the FMs. This is especially important when dealing with sensitive or personal data, where privacy must be carefully managed. By establishing clear access controls, data trustees protect the integrity of the data and allow the FMs to use it without compromising security or violating privacy laws.

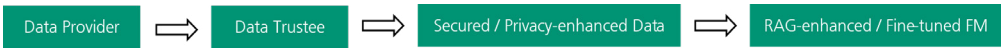

Figure 4. Illustration of the process by which data flows from the data provider to the data trustee, who ensures the secure and ethical handling of sensitive data before it is used in the development of an RAG-enhanced or fine-tuned FM.

Another important role of data trustees is to ensure privacy and compliance. With strict regulations such as the GDPR governing the use of personal data[40], compliance is paramount for FMs. Data trustees make sure that all data handling adheres to these legal frameworks, for example by providing centralized data anonymization. This ensures that personal data used to train FMs is handled ethically and legally, reducing risk for both data providers and model developers.

By acting as neutral intermediaries, data trustees increase trust between data owners. One of the major barriers to sharing high-quality data for FM training is the lack of trust between parties. Data owners are often reluctant to share valuable or sensitive data due to concerns about misuse or lack of control. Data trustees alleviate these concerns by managing data in a transparent and secure manner[41], giving data owners the confidence to contribute to FM training. This leads to richer and more diverse datasets that improve the quality and robustness of FMs. They also promote ethical and transparent development of FMs. By overseeing the data exchange process, they ensure that all steps are transparent and accountable, allowing data owners to understand how their data is being used. This contributes to the development of FMs that are not only technically proficient, but also ethically responsible. At a time when FMs are being scrutinized for bias and fairness, data trustees play a key role in ensuring that models are trained on diverse and ethically sourced data.

In addition to fostering trust, data trustees increase opportunities for collaboration. They provide a secure infrastructure for data exchange between different organizations, allowing stakeholders from different industries to work together without compromising data sovereignty. This is particularly important for industries that require high levels of privacy and security, such as healthcare and finance. By enabling the secure sharing of data across organizations, data trusts help FMs access industry-specific data that is often underutilized in FM training. This collaboration fosters innovation and strengthens the capabilities of FMs, making them more adaptable to specialized tasks.

4. Using Foundation Models and Data Spaces in an Industrial Context: Towards an R&D Agenda

As shown in this paper, the combination of data spaces and FMs offers many options that can open up new value chains for companies or improve existing ones. However, many aspects have not yet been fully clarified and need additional elaboration by research and industry. The following research questions require further investigation:

- What is the added value of extending FMs with data spaces?
- Does the use of data spaces in RAG or fine-tuning lead to more accurate answers?
- What are potential application scenarios for extended FMs?
- What is a reference architecture for integrating data spaces and FMs?
- Can data spaces address the current problems of FMs?
- How can the accuracy and quality of data from data spaces be measured and guaranteed?
- What is the usability (e.g., response time) of FMs extended with data spaces (RAG vs. fine-tuning)?
- What are the most effective methods for determining which information should be included in the context window of a query in a language model?
- Which techniques can be applied to select relevant data from a data space for fine-tuning a FM?
- How to ensure that the data usage policies from data spaces are also adhered to in FMs?

Fraunhofer ISST is committed to transforming research into industrial innovation. Industry leaders are invited to collaborate on advancing the integration of data spaces and foundation models to unlock new opportunities, overcome key challenges, and develop cutting-edge, trustworthy AI solutions. Partnering on this research agenda will ensure sustainable success and measurable value for all industries.

References

1 Gröger, C. (2021) *There is no AI without Data.* Communications of the ACM, 64(11), pp. 98-108.

2 Legner, C., Pentek, T., & Otto, B. (2020) *Accumulating Design Knowledge with Reference Models: Insights From 12 Years' Research into Data Management.* Journal of the Association for Information Systems, 21(3), p. 2.

3 Yee, L., Chui, M., Roberts, R., & Issler, M. (2024) *McKinsey Technology Trends Outlook 2024* [online]. McKinsey Digital. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-top-trends-in-tech#tech-trends-2024 [Accessed: 3 March 2025].

4 Banh, L. & Strobel, G. (2023) *Generative Artificial Intelligence.* Electronic Markets, 33(1), p. 63.

5 Dizikes, P. (2024) *Daron Acemoglu: What Do We Know About Economics and AI* [online]. MIT Economics. https://economics.mit.edu/news/daron-acemoglu-what-do-we-know-about-economics-ai [Accessed: 20 March 2025].

6 Gröger, C. (2021) *There is no AI without Data.* Communications of the ACM, 64(11), pp. 98-108.

7 Gröger, C. (2021) *There is no AI without Data.* Communications of the ACM, 64(11), pp. 98-108.

8 Amazon Web Services (n.d.) *What Are Foundation Models?* [online]. Amazon Web Services. https://aws.amazon.com/what-is/foundation-models/?nc1=h_ls [Accessed: 21 March 2025].

9 OpenAI (n.d.). *OpenAI* [online]. OpenAI. https://openai.com/ [Accessed: 21 March 2025].

10 Google Gemini (n.d.). *Gemini* [online]. Alphabet. https://gemini.google.com/ [Accessed: 21 March 2025]

11 Llama (n.d.). *Llama* [online]. Meta. https://www.llama.com/ [Accessed: 21 March 2025].

12 OpenGPT-X (n.d.) *Teuken 7B Instruct: Multilingual, Open Source Models for Europe – Instruction-Tuned and Trained in All 24 EU Languages* [online]. Akademie für Künstliche Intelligenz. https://opengpt-x.de/models/teuken-7b-de/ [Accessed: 21 March 2025].

13 Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024) *Generative AI.* Business & Information Systems Engineering, 66(1), pp. 111-126.

14 Manning, C. D., Raghavan, P., & Schütze, H. (2008) *Tokenization* in: Manning, C. D., Raghavan, P., & Schütze, H. (eds.) *Introduction to Information Retrieval* [online]. Cambridge University Press. https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html [Accessed: 20 March 2025].

15 Brown, T. B. (2020) *Language Models are Few-Shot Learners.* arXiv preprint arXiv:2005.14165

16 Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019) *Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding* in: Burstein, J., Doran, C., & Solorio, T. [eds.] *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* Association for Computational Linguistics, pp. 4171-4186.

17 Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019) *Fine-Tuning Language Models from Human Preferences.* arXiv preprint arXiv:1909.08593

18 Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., ... & Pentland, S. (2024) *Consent in Crisis: The Rapid Decline of the AI Data Commons.* arXiv preprint arXiv:2407.14933.

19 Amazon Web Services (n.d.) *What Are Foundation Models?* [online]. Amazon Web Services. https://aws.amazon.com/what-is/foundation-models/?nc1=h_ls [Accessed: 21 March 2025].

20 Otto, B. (2022) *The Evolution of Data Spaces in Designing Data Spaces: The Ecosystem Approach to Competitive Advantage.* Cham: Springer International Publishing, pp. 3-15.

21 Möller, F., Jussen, I., Springer, V., Gieß, A., Schweihoff, J. C., Gelhaar, J., ... & Otto, B. (2024) *Industrial Data Ecosystems and Data Spaces.* Electronic Markets, 34(1), p. 41.

22 European Committee for Standardization (2024) *Trusted Data Transaction* [online]. European Committee for Standardization. https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/RI/2024/cwa18125_2024.pdf [Accessed: 7 March 2025].

23 Data Spaces Support Centre (n.d.) *Introduction - Key Concepts of Data Spaces* [online]. Data Spaces Support Centre. https://dssc.eu/space/bv15e/766061351 [Accessed: 3 March 2025].

24 International Data Spaces Association (n.d.) *Dataspace Protocol 2024-1* [online]. International Data Spaces Association. https://docs.internationaldataspaces.org/ids-knowledgebase/dataspace-protocol [Accessed: 7 March 2025].

25 EDC (2025) *Data-Sharing at Scale* [online]. Eclipse Foundation. https://eclipse-edc.github.io/ [Accessed: 3 March 2025].

26 Gaia-X European Association for Data and Cloud AISBL (2025) *Together Towards a Federated and Secure Data Infrastructure* [online]. Gaia-X European Association for Data and Cloud AISBL. https://www.gaia-x.eu/ [Accessed: 3 March 2025].

27 Catav, A., Miara, R., Giloh, I., Cordeiro, N. & Ingber, A. (2024) *RAG Makes LLMs Better and Equal* [online]. Pinecone. https://www.pinecone.io/blog/rag-study/ [Accessed: 21 March 2025].

28 DataScientest (2023) *Word2vec: NLP & Word Embedding* [online]. DataScientest. https://datascientest.com/de/word2vec [Accessed: 21 March 2025] (in German).

29 Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv preprint arXiv:1810.04805

30 Zhao, S., Yang, Y., Wang, Z., He, Z., Qiu, L. K., & Qiu, L. (2024) *Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make Your LLMs Use External Data More Wisely.* arXiv preprint arXiv:2409.14924

31 Pinecone (o.d.) *Build Knowledgeable Ai. The Vector Database for Scale in Production* [online]. Pinecone https://www.pinecone.io/ [Accessed: 21 March 2025].

32 Meta (n.d.) *Tools: Faiss* [online]. Meta. https://ai.meta.com/tools/faiss/ [Accessed: 21 March 2025].

33 Yadav, D., Para, H., & Selvakumar, P. (2023) *Unleashing the Power of Large Language Model, Textual Embeddings, and Knowledge Graphs for Advanced Information Retrieval.* 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET), pp. 1-5.

34 Zhao, S., Yang, Y., Wang, Z., He, Z., Qiu, L. K., & Qiu, L. (2024) *Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make Your LLMs Use External Data More Wisely.* arXiv preprint arXiv:2409.14924

35 Yu, T., Xu, A., & Akkiraju, R. (2024) *In Defense of RAG in the Era of Long-Context Language Models.* arXiv preprint arXiv:2409.01666

36 Lu, W., Luu, R. K., & Buehler, M. J. (2024) *Fine-tuning Large Language Models for Domain Adaptation: Exploration of Training Strategies, Scaling, Model Merging and Synergistic Capabilities.* arXiv preprint arXiv:2409.03444

37 Yun, J., Sohn, J. E., & Kyeong, S. (2023) *Fine-tuning Pretrained Language Models to Enhance Dialogue Summarization in Customer Service Centers.* International Conference on AI in Finance. https://doi.org/10.1145/3604237.3626838

38 Zou, J., Zhou, M., Li, T., Han, S., & Zhang, D. (2024) *PromptIntern: Saving Inference Costs by Internalizing Recurrent Prompt during Large Language Model Fine-tuning.* arXiv:2407.02211. https://doi.org/10.48550/arXiv.2407.02211

39 Bundesdruckerei (2025) *Data Trustee. Data Trustee Platform with a Trust Center Service on Demand* [online]. Bundesdruckerei. https://www.bundesdruckerei-gmbh.de/en/solutions/data-trustee [Accessed: 7 March 2025].

40 Delacroix, S. & Montgomery, J. (2020) *Data Trusts and the EU Data Strategy* [online]. Data Trusts Initiative. https://datatrusts.uk/blogs/data-trusts-and-the-eu-data-strategy [Accessed: 21 March 2025].

41 Feth, D., Rauch, B., Krohmer, D., von Albedyll, J. & Barreto Villela, K. (2022) *Datentreuhänder – Begriffliche Einordnung und Definition (Teil 1)* [online]. Fraunhofer IESE Blog. Available at: https://www.iese.fraunhofer.de/blog/datentreuhaender-definition/ [Accessed: 21 March 2025] (in German).

Image Credits

©Lee – AdobeStock, title

Contact

Dr.-Ing. Marcel Altendeitering
Head of Department Mobility & Smart Cities

marcel.altendeitering@isst.fraunhofer.de
+49 231 97677-461

Fraunhofer-Institute for Software and Systems Engineering ISST
Speicherstraße 6
D-44147 Dortmund
Germany

www.isst.fraunhofer.de