

UNIVERSITÄT  
BAYREUTH

Bayreuther Arbeitspapiere zur Wirtschaftsinformatik  
*Bayreuth Reports on Information Systems Management*

Lars Böcking, Anne Michaelis, Bastian Schäfermeier, André Baier,  
Niklas Kühl, Marc-Fabian Körner, Lars Nolting

**Generative Artificial Intelligence  
in the Energy Sector**



No. 71  
April 2024  
ISSN 1864-9300





# Generative Artificial Intelligence in the Energy Sector

---

**Disclaimer**

This study was created by the Fraunhofer Institute for Applied Information Technology FIT (Fraunhofer FIT), the Fraunhofer Institute for Energy Economics and System Technology (Fraunhofer IEE), and TenneT TSO GmbH to the best of their knowledge and with due diligence.

The Fraunhofer FIT, the Fraunhofer IEE, and TenneT TSO GmbH, their legal representatives, and/or auxiliary agents provide no warranty whatsoever that the content of this thesis paper is verified, complete, usable for specific purposes, or otherwise free from errors. Use of this thesis paper is entirely at your own risk. Under no circumstances do the Fraunhofer FIT, the Fraunhofer IEE, and TenneT TSO GmbH, their legal representatives, and/or auxiliary agents assume responsibility for any damage that directly or indirectly results from the use of this thesis paper.

**Recommended citation style**

Böcking, Lars; Michaelis, Anne; Schäfermeier, Bastian; Baier, André; Kühl, Niklas; Körner, Marc-Fabian; Nolting, Lars (2024). Generative Artificial Intelligence in the Energy Sector. Published by Fraunhofer FIT, Fraunhofer IEE and TenneT TSO GmbH. Bayreuth. [https://doi.org/10.15495/EPub\\_UBT\\_00007674](https://doi.org/10.15495/EPub_UBT_00007674).

## Authors



Lars Böcking

Fraunhofer FIT



Anne Michaelis

Fraunhofer FIT

Dr. Bastian  
Schäfermeier

Fraunhofer IEE



André Baier

Fraunhofer IEE

Prof. Dr. Niklas  
Kühl

Fraunhofer FIT

Dr. Marc-Fabian  
Körner

Fraunhofer FIT



Dr. Lars Nolting

TenneT TSO GmbH

The Fraunhofer Cluster of Excellence Integrated Energy Systems (CINES), consisting of the Fraunhofer FIT and the Fraunhofer IEE, among others, has earned an excellent international reputation in research due to its interdisciplinary subject areas. Their research areas energy and energy technology not only deal with the future-oriented topic of energy from a natural and engineering science angle but also in terms of socio-political, economic and legal aspects. Fraunhofer CINES has proven expertise in the ability to combine methodological know-how at the highest scientific level with a customer-focused and solution-oriented way of working, which is our distinctive feature.

As part of an internal initiative, the Fraunhofer Gesellschaft pursues the goal of getting carbon-neutral by 2030. The transmission system operators have already committed themselves to reducing their carbon footprint while continuing to guarantee a very high security of supply and offering fair energy prices.

TenneT is among the leading electricity grid operators in Europe and is committed to ensuring a highly secure and reliable energy supply 24 hours a day, 365 days a year.

As the first cross-border transmission system operator, TenneT is planning, building, and operating an almost 25,000 kilometer-long high and extra-high voltage grid in the Netherlands and large parts of Germany and supports the European energy market through 16 interconnectors to neighboring countries. TenneT belongs to the largest investors in the European energy transition and contributes to shaping a sustainable, reliable, and affordable energy supply system fit for the future, in which digitalization plays an essential role. With about 7,400 internal and external employees and corporate values of responsibility, courage, and networking, TenneT ensures that over 43 million end users in Europe are supplied with stable power that they can rely on every day.

# Table of Contents

---

<b>Foreword</b>	<b>5</b>
<b>Introduction</b>	<b>6</b>
Definition of Generative Artificial Intelligence	6
Generative Artificial Intelligence vs. Artificial Intelligence	6
Motivation of Topic	8
Overview of GenAI Methods	9
<b>State-of-the-Art Analysis of Generative Artificial Intelligence</b>	<b>10</b>
General concepts of Generative Artificial Intelligence	10
Generative Adversarial Networks	10
(Variational) Autoencoder	12
Diffusion Models	13
Foundation Models	14
Concepts in Large Language Models	16
Pre-training	16
Fine-Tuning	17
Reinforcement Learning from Human Feedback	18
Connect Large Language Models to Tools	19
Large Language Model Agents	21
Large Language Model Information Retrieval	23
Large Language Model Prompt Engineering	24
<b>GenAI-based Use Cases for TenneT</b>	<b>26</b>
Use Case 1: Public Affairs & Communication Support	27
Use Case 2: Interactive Corporate Knowledge Base	29
Use Case 3: Energy Sector Professor	31
Use Case 4: Grid Maintenance Agent	33
<b>Conclusion &amp; Recommendations for Action</b>	<b>36</b>
<b>References</b>	<b>38</b>



## Foreword

---

In the domain of generative artificial intelligence (GenAI), current trajectories suggest a **paradigmatic transition** across the industrial spectrum. Quantitative forecasts propose that GenAI technologies could catalyze an augmentation of the global Gross Domestic Product (GDP) by an estimated 7% over the next decade [1]. This economic expansion is projected to have an annual fiscal impact on the global economy, ranging from \$2.6 to \$4.4 trillion [2]. AI-driven innovation is becoming increasingly critical for maintaining market leadership and operational efficacy. Corporations now need to assess GenAI use cases to gain competitive advantages in a rapidly evolving digital landscape.

The transformative power of GenAI extends beyond mere economic metrics. We will see a **shift in the functional roles** across industries. Notably, functions such as marketing, sales, and product R&D are projected to experience a substantial impact, measured both in dollar terms and as a percentage of functional spending. Concretely, marketing is projected to absorb a global impact of upwards of \$450 billion, sales are forecasted to experience an impact of nearly \$480 billion, and product R&D is estimated at approximately \$320 billion [1]. With an even higher impact ratio compared to the functional spend, software engineering, and customer operations will undergo an even larger GenAI transformation. Cumulatively, these segments are expected to assimilate approximately three-quarters of the aggregate annual financial impact engendered by GenAI applications. These potentials for various organizational functions highlight the strategic implications GenAI can have for companies across industries.

**The energy sector** stands out to be facilitated by GenAI applications. Projections situate the potential impact on productivity within a range of 1.0–1.6% [1]. This translates to an anticipated global economic valuation impact spanning from \$150 to \$240 billion [1]. The implications of these projections are multifaceted, encompassing not only quantifiable finan-

cial gains but also the strategic enhancement of operational efficiencies and the adoption of innovative energy management practices. The integration of GenAI is thus poised to drive forward the sector's capabilities in predictive analytics, grid management, and the seamless assimilation of renewable energy sources, thereby fortifying the foundational robustness of critical energy infrastructures. This is highly relevant against the backdrop of (1) increasing complexity in the energy sector and (2) increasing pressure to deliver due to the energy transition.

### **This study is structured as follows:**

The study is divided into two parts. The first part is intended for the energy sector in general, while the second part presents TenneT-specific use cases.

The initial sections dissect GenAI concepts from traditional Machine Learning concepts and provide more extensive background on the topic at hand. The section State-of-the-Art Analysis of GenAI is split into three main parts. First, various GenAI models such as Generative Adversarial Networks (GANs), (Variational) Autoencoders, and Diffusion Models are discussed. Secondly, this section focuses on Large Language Models and various adaptations and extensions. The section is finalized by a technology map setting the discussed concepts into relation. The next main section, GenAI-based use cases, is TenneT-specific and explores expert-picked use cases in the energy sector in greater depth. Besides a detailed outlined solution, an extensive risk and potential analysis is provided, diving deeper into evaluation dimensions such as the expected value, the maturity of the technology as well as the data availability and compliance. The final section provides a conclusion and concrete recommendations for action to make the best of the GenAI revolution in the energy sector.

# Introduction

## Definition of Generative Artificial Intelligence

**Artificial intelligence** (AI) is an umbrella term for methods and tools in computer science that simulate intelligent behavior, such as reasoning, perception, problem-solving, learning, or communication. **Generative AI** refers to a branch of artificial intelligence in which models are capable of generating new data such as text, images, or sound (e.g., speech).

Prominent examples of such models are large language models (LLMs) such as ChatGPT and widely used tools for image synthesis, such as Stable Diffusion or Dall-E. In audio synthesis, Google’s WaveNet and WaveRNN are used to generate natural-sounding speech synthesis for tools such as Google Assistant or Maps Navigation.

For a more mathematically oriented definition, GenAI models are designed to learn a data distribution  $p(x)$  based on a set of training data examples  $x \in X$  and allow for sampling from this distribution to generate new data.

## Generative Artificial Intelligence vs. Artificial Intelligence

Generative models can be distinguished from **discriminative models**, which learn to discriminate between different classes of data based on annotated training examples. Mathematically, such models learn the distribution  $p(y|x)$ , i.e., the distribution of class labels  $y \in Y$  conditioned on data samples  $x \in X$ . For this, discriminative models typically require labeled data examples for their training. For example, a computer vision model that is trained to distinguish between photos of different kinds of animals requires a set of training images that are each labeled with the name of the depicted animal. Figure 1 (left) depicts generative AI in the context of discriminative AI.

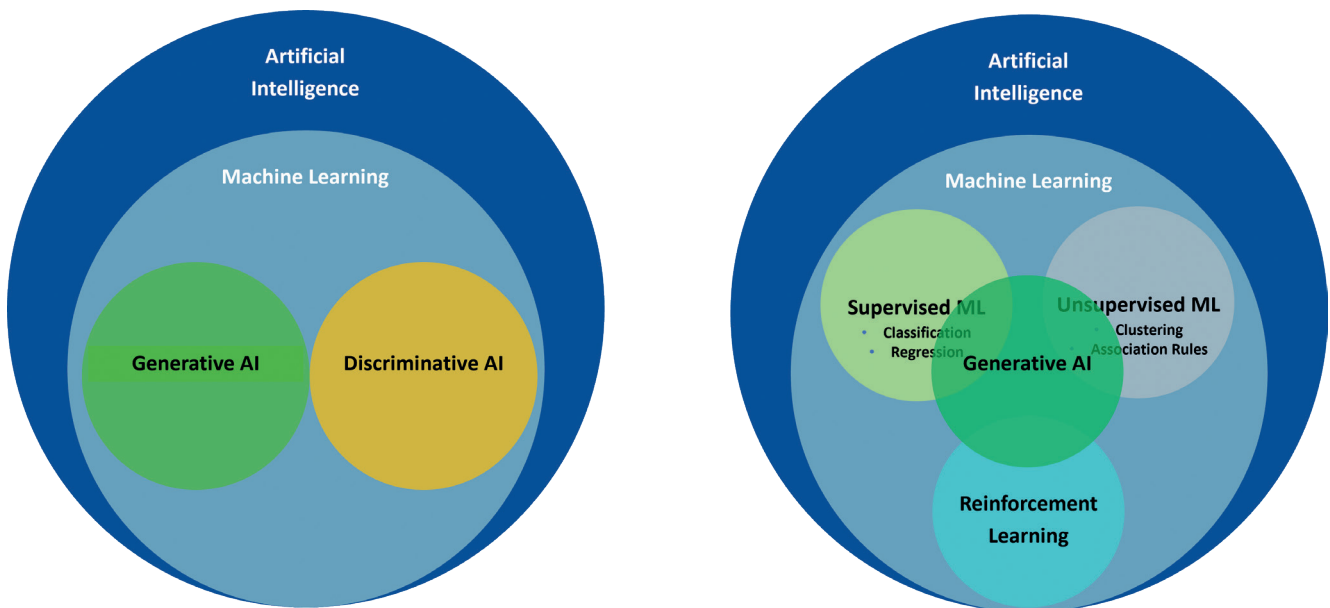


Figure 1: Left: Venn diagram of GenAI in the context of discriminative AI. Right: GenAI in the context of other AI-related concepts.

Preparing labeled data often requires considerable manual effort. In contrast to this, many generative models, e.g., for text generation, do not require curated and labeled training data but can be trained on large amounts of arbitrary, unlabeled data such as texts or images from the world wide web. This potential of leveraging much larger amounts of data makes them very powerful and capable of learning many details and variations within the data distribution.

Some further important terms and concepts from AI that are useful but not necessary for the comprehension of the present study and for putting GenAI into context of AI in general are described in the table below. To put the concepts into context, another Venn diagram is given in Figure 1 (right).

Table 1: Overview of AI-related terms and concepts.

<b>Machine Learning</b>	AI methods which involve some kind of “learning” process, e.g., learning from training data examples or through trial and error.
<b>Supervised Learning</b>	A subset of machine learning methods that learn from labeled examples. Important methods here are classification and regression. In classification, examples are labeled by their class (i.e., elements of a set $y \in Y$ ). In regression, they are labeled with a real number $y \in R$ . In the supervised learning process, a classification or regression model learns to predict the dependent variable (i.e., class or regressand).
<b>Unsupervised learning</b>	Subset of machine learning methods that learn from unlabeled data, e.g., based on patterns in the data or their structure. One important technique here is clustering, i.e., discovering groups of similar items in data. Clustering is similar to classification but based on the inherent properties and structure of the data rather than predefined class labels.
<b>Reinforcement Learning</b>	Machine Learning methods which learn to perform a series of actions within an environment to fulfill tasks based on received feedback (rewards or penalties). Reinforcement learning is used for tasks where the solution quality can be scored easily based on a “reward” function and where the path to the solution can be complex. Examples of applications include game-playing and LLM fine-tuning.
<b>(Artificial) Neural Networks</b>	<p>Neural networks are machine learning models that have been particularly successful in machine learning in recent years and are inspired by the structure of the human brain. They consist of “neurons” with activation functions and connections and are typically structured as sets of subsequent layers. Many kinds of neural networks and architectures exist, which can be used for different tasks such as classification, regression, and as generative models.</p> <p>Important architectures include the multilayer perceptron (MLP, typically used for classification or regression), convolutional neural network (CNN, typically used for images), recurrent neural networks (RNN), and Long Short-Term Memory (LSTM), which are both used for sequential data such as time series. For text generation, the transformer model has become the widely used standard. The architectures may also be used as building blocks, which can be combined, e.g., convolutions with an MLP head for image classification.</p>
<b>Deep Learning</b>	Refers to deep neural networks, i.e., artificial neural networks, which consist of many layers.



### Motivation of Topic

While GenAI has a long history in AI research, it became prominent and widely used in recent years, in large parts due to LLM applications such as **ChatGPT** and **image generation** applications such as Stable Diffusion. Another reason was strong advances in artificial neural networks in general, which have made them applicable and used in all kinds of areas, from autonomous driving to medicine.

It is foreseeable that these methods will continue to have strong influences on all sorts of areas. Among these areas, we see great potential for a transformation of the energy sector. We aim to contribute to this transformation through our study, which comprises a state-of-the-art analysis of GenAI with a focus on methods and possible applications in the energy sector. The first half of our study consists of two main parts: First, a description of general GenAI methods and, second, a more detailed overview on LLM-based methods and concepts. For each described approach, we outline possible use cases within the energy sector.

### Overview of GenAI Methods

In practice, GenAI-based solutions are often comprised of many different methods, data sources, and tools, which are combined into an overall system.

An overview of how methods are typically combined for LLMs is given in Figure 7. As the figure shows, state-of-the-art LLMs involve a variety of building blocks and steps. The LLM must be pre-trained on a large general text corpus to create a foundation model. In a next step, the model is often fine-tuned to a more specific domain or use case. To be usable as a chatbot and mitigate, e.g., toxic behavior, the model is aligned to follow human instructions through Reinforcement Learning from Human Feedback (RLHF). Once these model preparation steps are dealt with, the final model can be connected to external knowledge, either through a vector database or through tools such as search engines or Wikipedia. For connecting external tools, e.g., a Python interpreter that can execute program code

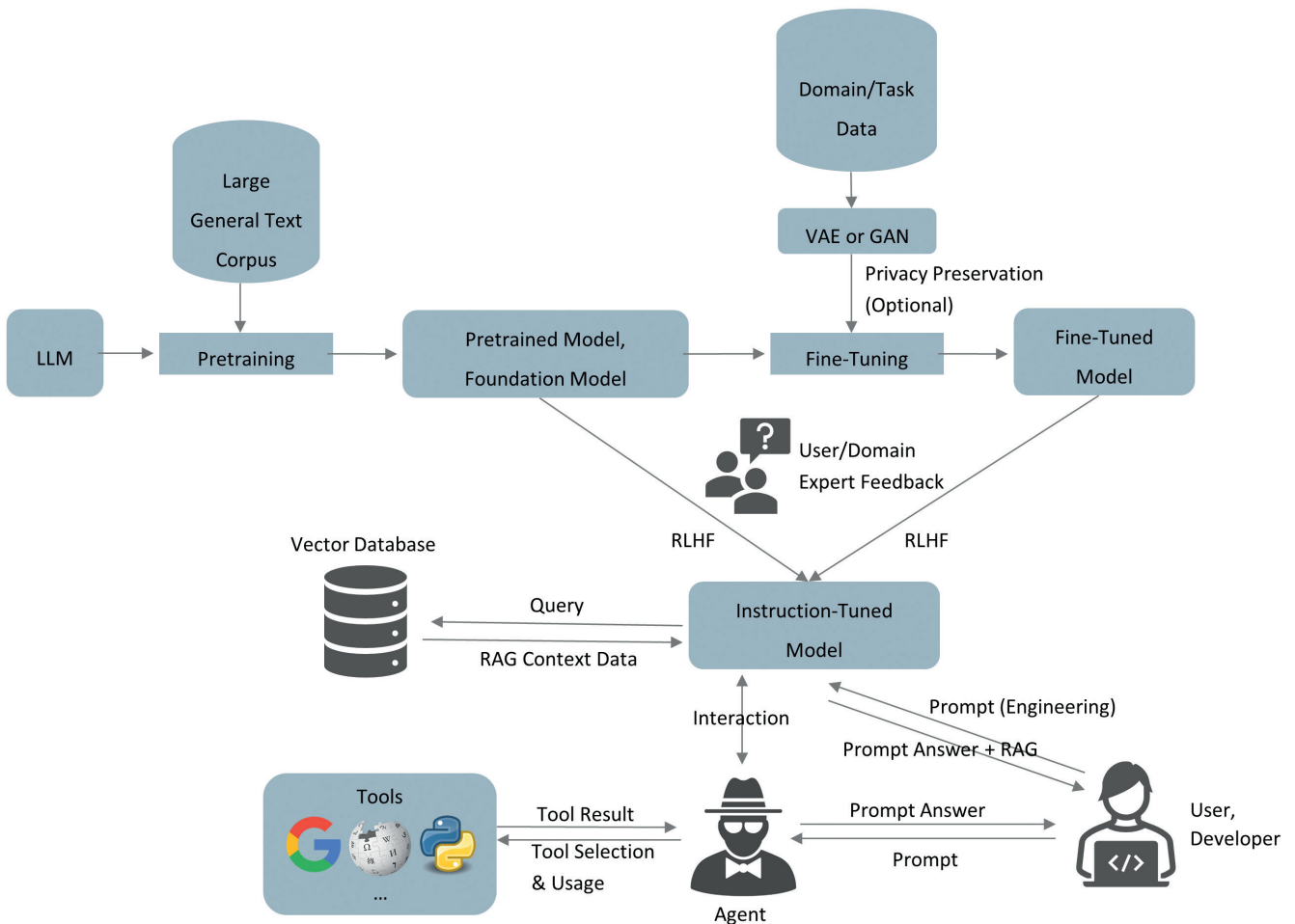


Figure 2: Overview of LLM methods and how they can be combined.

generated by the LLM, some kind of basic agent model must be implemented, which decides when to use such a tool based on the LLM's in- and outputs. The mentioned building blocks will be explained in more detail in this study.

Figure 8 gives a simplified view of how the aforementioned techniques are integrated with ChatGPT. GPT-4 was pre-trained on texts from books and articles, e.g., from Wikipedia, and websites. RLHF was used for the alignment step. ChatGPT allows for uploading and storing files of different formats. Uploaded files can be queried for contents through RAG. Integrated tools involve, e.g., web search, image generation

through Dall-E, and image interpretation through GPT-4 Vision, a multimodal text- and vision model that allows to create detailed descriptions of images. Since OpenAI is not open about all technical details of GPT4, it is unknown how, e.g., tool selection is performed in detail. Some kind of selection mechanism based on tool descriptions similar to an agent model is likely to be involved, e.g., for web search and image creation. Tools for handling uploaded files, e.g., extracting contents and storing them in a vector database, are probably hard-coded into the system.

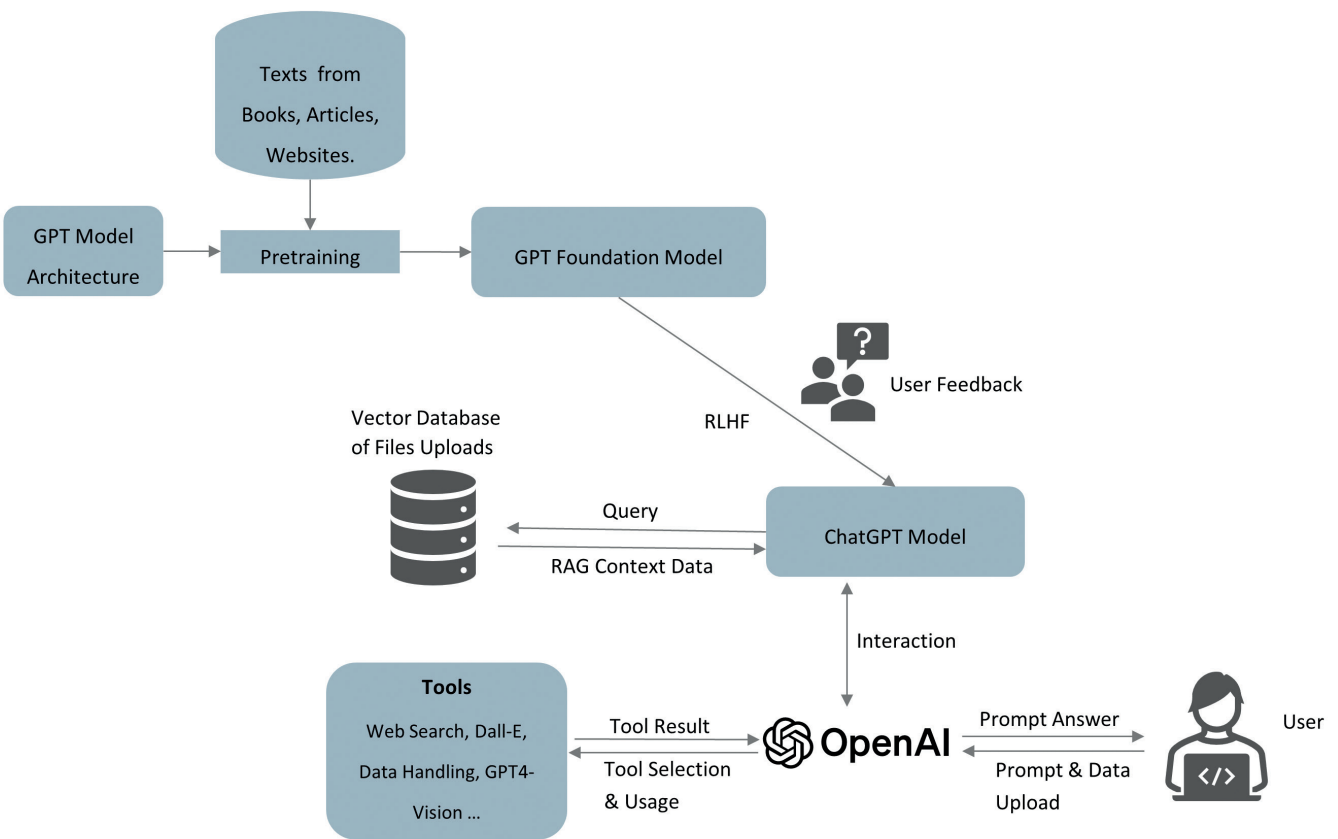


Figure 3: Overview of LLM methods integrated in ChatGPT.

# State-of-the-Art Analysis of Generative Artificial Intelligence

This section provides an in-depth analysis of GenAI and its recent advancements. The primary objective is to examine the state-of-the-art of GenAI by providing a comprehensive understanding of the complex mechanisms underlying GenAI and its multifaceted capabilities.

Following each technical background, the study explores the various fields of application where the specific GenAI technology can be useful in the energy sector. It presents illustrative use cases to demonstrate how GenAI optimizes processes, enhances operational efficiencies, and fosters innovation in the energy domain. Furthermore, each section assesses the potential benefits and risks of implementing GenAI in the energy sector.

The methodological framework of this analysis includes a literature review, which is a crucial aspect. The section aims to provide a comprehensive understanding of the prevailing trends, challenges, and opportunities in the field of GenAI, with a focus on applications in the energy sector. It draws on a diverse array of academic sources, industry reports, and empirical studies.

## General concepts of Generative Artificial Intelligence

This section outlines general concepts of GenAI, providing both intuitive overviews and detailed technical insights. In addition, each concept is examined for its inherent advantages and disadvantages, complemented by three distinct use cases that illustrate its practical importance:

1. **Generative Adversarial Networks (GANs)** employ a dual-network architecture to generate synthetic data samples through adversarial training.
2. **(Variational) Autoencoder** utilize an encoder-decoder framework to learn efficient data representations for tasks such as compression and generation.
3. **Diffusion Models** gradually transform patterns of random noise into structured data, such as images or text, by reversing a diffusion process.
4. **Foundation Models**, trained on diverse datasets and fine-tuned for specific tasks, exhibit strong performance across various applications without task-specific adaptation.

## Generative Adversarial Networks

**Generative Adversarial Networks (GAN [3])** are a class of generative models mostly used in **image synthesis**. GANs consist of two neural networks: A generator and discriminator network. The generator generates images based on input noise and, optionally, input conditions. The discriminator is a classifier which distinguishes between real data examples and “fake” examples generated by the generator.

In the training process, both networks compete against each other: The generator is trained to generate fake examples that are as realistic as possible, i.e., which the discriminator is not able to distinguish from real examples. The discriminator, on the other hand, is trained to recognize the fake examples and distinguish them from real examples. Both networks improve more and more on their task, such that, after the training procedure, the generator can produce realistic data samples.



Figure 4: Left = CycleGAN Architecture. Right = CycleGAN style transfer example application.

One important GAN variant includes the **DCGAN**, a GAN architecture based on a convolutional neural network typically used for image generation. A problem with “vanilla” DCGANs is that they are difficult to train. The discriminator and generator must be well-balanced. If the discriminator can recognize all fake samples or no fake samples at all, the generator cannot learn how to generate convincing samples. Another problem is “mode collapse”, i.e., the generator learns to generate some convincing samples but repeats these instead of learning the full data distribution. To mitigate these issues, **Wasserstein GANs** (WGANs [4], [5]) are trained to minimize the so-called Wasserstein Loss between the training data and the generated data distribution, which measures their similarity and, therefore, encourages the GAN to generate more diverse samples.

Another important variant are **CycleGANs** [6], which consist of two GANs which translate data examples between two domains (“style transfer”) and are typically used as a kind of

“filter” for images. Some examples of this include converting photorealistic images into the painting style of a specific artist and vice versa (see Fig. 4) or converting satellite photos into a schematic map representation. In autonomous driving, they are used to generate more realistic training videos based on rendered images from a 3D simulation.

Finally, **conditional GANs** incorporate conditions to control the output of GANs. Conditions may be a desired output class or properties of the sample to generate [7], but also other images [8]. Further interesting applications of GANs are face synthesis (e.g., <https://thispersondoesnotexist.com/> [9]), image denoising, and image upscaling. In the energy sector, they are, for example, used by Fraunhofer IEE to synthesize time series, such as power curves. The generated time series for household power consumption are used for electrical load estimation in power networks.

Generative Adversarial Network use cases in the energy sector	
<b>Remote sensing through satellite images</b>	
<ul style="list-style-type: none"> <li>■ Satellite images can be converted to a more “schematic” representation through CycleGANs, e.g., to extract properties of sites (land use type)</li> <li>■ Application to network planning</li> </ul>	
<b>Improve maintenance images</b>	
<ul style="list-style-type: none"> <li>■ Use GANs or CycleGANs to upscale or denoise images</li> <li>■ E.g., for power line inspection through camera-equipped drones or helicopters</li> </ul>	
<b>State estimation of power networks</b>	
<ul style="list-style-type: none"> <li>■ Generate synthetic power load time series for consumers</li> <li>■ Based on this, perform state estimation, e.g., through pandapower</li> </ul>	
Potential	Risk
<p>low potential                      high potential</p>	<p>low risk                              high risk</p>
<ul style="list-style-type: none"> <li>■ (Conditional) GANs work well for <b>generating images with given categories or features</b></li> <li>■ There are certain GAN architectures for <b>style transfer</b> (“CycleGANs”), <b>image denoising</b> and <b>upscaling</b></li> </ul>	<ul style="list-style-type: none"> <li>■ “Vanilla” GANs are extremely <b>difficult to train</b></li> <li>■ Suffer from <b>“mode collapse”</b>, i.e., generating some, convincing images while failing to capture the full distribution of the training data (Wasserstein GANs improve upon this)</li> </ul>



**(Variational) Autoencoder**

The concept of Autoencoders (AEs) was established as a framework for the efficient encoding of input data into succinct representations [10], [11], [12]. The quintessence of an AE lies in its capacity to compress data through an encoder function  $e_{\theta}(x)$  and subsequently reconstruct the input via a decoder function  $d_{\phi}(z)$ , striving to minimize a loss function, typically the mean squared error  $\mathcal{L}(x, \hat{x}) = ||x - \hat{x}||_2$ , where  $x$  is the original input and  $\hat{x}$  is the reconstructed output. The AE stores information in the latent representation  $z$ , a deterministic data encoding as fixed vectors [13].

Variational Autoencoders (VAEs) are as a sophisticated extension of AEs, transcending the deterministic confines by introducing a stochastic element to the encoding process, shown in Figure 5. Unlike their deterministic counterparts, VAEs conceptualize the latent space as a probabilistic distribution, which enables the generation of new data instances through the sampling of latent variables  $z = e_{\theta}(x)$  from a defined distribution  $\mathcal{N}(\mu_x, \sigma_x^2)$  [14], [15], [16].

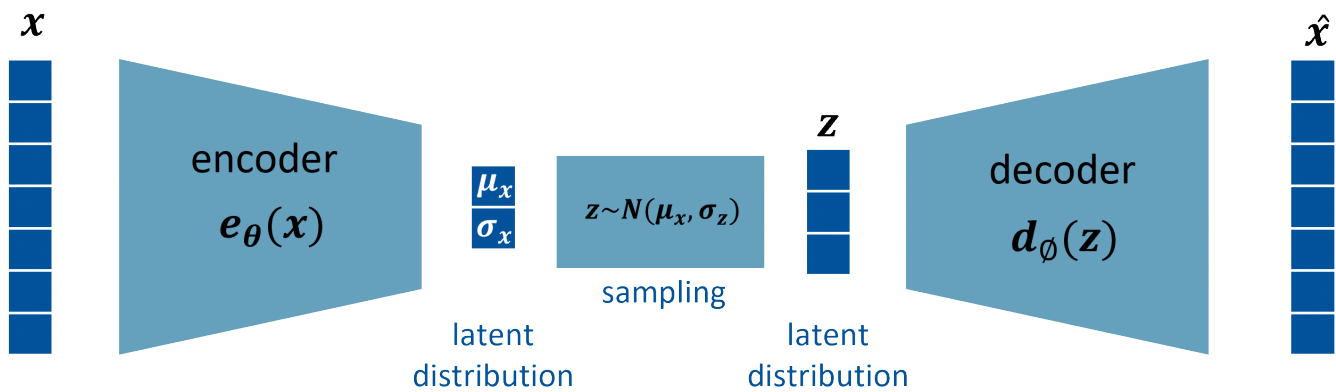




Figure 5: Concept Variation-Autoencoder

The generative ability of Variational Autoencoders (VAEs) is highly dependent on the parameterization of the latent space. One needs to balance two **competing objectives**: the reconstruction quality, which ensures that the output closely resembles the original data, and the regularization of the latent space, which enforces a degree of smoothness and continuity. The balance between fidelity to the input data and deviation from a prior distribution is achieved through the loss function, which consists of the reconstruction loss and the *Kullback-Leibler divergence*.

VAEs have been extended and adapted to meet the specific demands of diverse data types and application requirements. **Temporal difference VAEs** [17], for instance, have been engineered to process time-series data effectively, which is particularly relevant for industries like energy where consumption patterns unfold over time. Further developments have led to the creation of **VAEs with interpretable, factorized latent representations**, particularly for image data. In the realm of energy provision, such advancements could facilitate the analysis of drone or satellite imagery [12]. The imperative for **algorithmic fairness**, especially when processing sensitive factors, has led to adaptations that include fairness constraints in Autoencoder architectures [18].

Variational Autoencoder use cases in the energy sector	
<p><b>Anomaly detection</b></p> <ul style="list-style-type: none"> <li>Train an AE to reconstruct energy load patterns, input real patterns, and identify mismatches that represent anomalies.</li> <li>Derive predictive maintenance measures for energy infrastructure.</li> </ul>	
<p><b>Generate energy consumption scenarios</b></p> <ul style="list-style-type: none"> <li>Train VAE on the historical energy consumption data to learn a latent representation of energy patterns.</li> <li>Generate new synthetic data points that represent plausible scenarios.</li> </ul>	
<p><b>Denoising real-world data stream</b></p> <ul style="list-style-type: none"> <li>Use VAEs to identify underlying trends and patterns in noisy, high-dimensional data, such as individual household power consumption.</li> <li>The anonymization of data can also be a beneficial side effect.</li> </ul>	
Potential	Risk
 <p>low potential                      high potential</p>	 <p>low risk                              high risk</p>
<ul style="list-style-type: none"> <li>VAEs are adept at capturing the underlying distribution of data, making them highly effective for anomaly detection in energy load patterns. This allows the model to precisely <b>learn what constitutes normal operation</b> and highlight deviations that signal maintenance needs.</li> <li>VAEs can <b>generate new data points</b> that are not mere replicas of the input data but rather variations thereof, e.g., providing high-quality and diverse energy consumption scenarios for planning and stress testing.</li> </ul>	<ul style="list-style-type: none"> <li>Training VAEs requires <b>careful tuning of the loss function</b>, especially the balance between the reconstruction loss and the Kullback-Leibler divergence. This process can be complex and time-consuming.</li> <li>Although the scenarios generated by VAEs are realistic, they may still require <b>expert interpretation to translate into actionable insights</b>. This is because the connection between latent variables and physical phenomena may not always be clear.</li> </ul>

**Diffusion Models**

Diffusion models [19], [20], similar to GANs and VAEs, are used in image synthesis. The basic principle of these methods is to model a noising and denoising process of an image, which are disassembled into several time steps. In the forward diffusion process, noise is added to the training images until gradually, only noise is left. In the reverse diffusion process, the model

learns to gradually remove the noise to generate output images. Typically, diffusion models incorporate text to condition the output. This allows to **generate images based on text prompts**. Some well-known models in this area include *Stable Diffusion*, *OpenAI's Dall-E [21]*, and *Midjourney*.

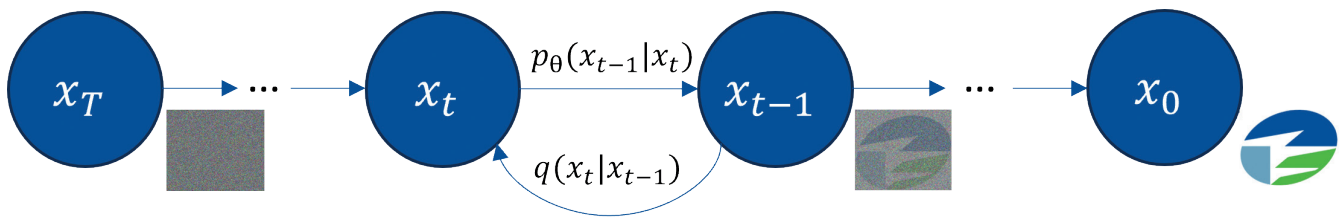


Figure 6: Forward and reverse diffusion process. Source: [19]

Diffusion Model use cases in the energy sector	
<b>Generate illustrations for marketing</b>	
<ul style="list-style-type: none"> <li>Models like Stable Diffusion can be used to create copyright-free illustrations</li> <li>E.g., for website articles, brochures or presentations</li> </ul>	
<b>Generate maintenance training data</b>	
<ul style="list-style-type: none"> <li>Models for Image Generation can be used to create or augment training data for computer vision tasks</li> <li>E.g., automated power line inspection</li> </ul>	
<b>Generate power load time series</b>	
<ul style="list-style-type: none"> <li>Diffusion models are also applicable to other kind of data, e.g., time series</li> <li>Create synthetic power load time series data</li> <li>Advantage: Reduced legal issues in further processing of synthetic data</li> </ul>	
Potential	Risk
<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <span>■</span> <span>■</span> <span>■</span> <span>■</span> <span>□</span>                      low potential                 </div> <div style="text-align: center;"> <span>□</span> <span>□</span> <span>□</span> <span>□</span> <span>□</span>                      high potential                 </div> </div>	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <span>■</span> <span>■</span> <span>□</span> <span>□</span> <span>□</span>                      low risk                 </div> <div style="text-align: center;"> <span>□</span> <span>□</span> <span>□</span> <span>□</span> <span>□</span>                      high risk                 </div> </div>
<ul style="list-style-type: none"> <li>Diffusion models are exceptionally good at <b>generating images from text prompts</b>.</li> <li>It can also be used for <b>time series and audio synthesis</b> (although not very common).</li> </ul>	<ul style="list-style-type: none"> <li>Diffusion models <b>require careful prompt engineering</b> to generate good images.</li> <li>Parts of text prompts may be ignored by a model or not work as expected.</li> <li>Due to random elements, it requires some <b>trial and error</b> for tuning results.</li> <li>Potential legal/copyright issues when it is unclear where training data came from.</li> </ul>



### Foundation Models

Foundation models are at the forefront of the GenAI field. They are a type of deep learning model that has been trained on **extensive corpora of unstructured and unlabeled data**. Examples of such models include GPT-4, PaLM, DALL-E 2, and Stable Diffusion. These models have demonstrated remarkable versatility and can be used for a wide range of tasks straight out of the box or fine-tuned for specific applications.

The term foundation model was originally coined by the *Stanford Center for Research on Foundation Models (CRFM)*, as "A foundation model is any model that is trained **on broad data (generally using self-supervision at scale)** that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" [22]. For companies operating within the data-rich domain of energy provision, the utility of foundation models is particularly promising.

These models are not only capable of learning vast amounts of unlabeled data but can also process data of **multiple modalities**. Even though many of today’s discussed applications focus on the processing of natural language, the CRFM group “understand foundation models as a **general paradigm of AI**, rather than specific to NLP in any way” [22]. Foundation models can process written text like in the classical Natural Language Processing domain, as well as images similar to the computer vision domain or tabular data on a competitive performance level to classical machine learning algorithms tailored for such applications. Recent advancements in the field, such as Gemini by Google Deepmind in December 2023, specifically highlight their abilities in multimodal benchmarks [23].

The Transformer architecture is at the core of many contemporary foundation models. It uses a self-attention mechanism to evaluate input sequences in their entirety, thereby capturing long-range dependencies and contextual nuances more effectively than previous model generations [24]. This architectural choice enables foundation models to excel in a range of natural language processing tasks, including translation, summarization, and complex text generation.

Foundation model use cases in the energy sector	
<p><b>Onboard new employees on projects</b></p> <ul style="list-style-type: none"> <li>Screen various data sources such as initial concept papers, midterm presentations, meeting recordings, and code implementation.</li> <li>Summarize the current project state for newly onboarded team members.</li> </ul>	
<p><b>Company internal knowledge base</b></p> <ul style="list-style-type: none"> <li>Have a chatbot that can screen various unstructured data formats.</li> <li>Employees can search for company experience in e.g., “redispatch 2.0 processes”.</li> </ul>	
<p><b>Cross-border energy market analysis</b></p> <ul style="list-style-type: none"> <li>Track energy prices and developments in real time and combine them with data on regulatory changes.</li> <li>Analyze cross-border energy consumption patterns and identify any irregularities and concept shifts</li> </ul>	
Potential	Risk
 <p>low potential <span style="margin-left: 150px;">high potential</span></p>	 <p>low risk <span style="margin-left: 100px;">high risk</span></p>
<ul style="list-style-type: none"> <li>Foundation models, due to the training on diverse and extensive datasets, <b>capture a wide breadth of knowledge</b>. This means a single model could understand terminology from energy markets, technical engineering language, and even regulatory compliance details.</li> <li>These models can be quickly adapted to perform a variety of tasks and modalities since initial pattern recognition has already been trained. This means <b>fast deployment across different departments</b>—from legal compliance inquiries to technical project summaries and visual maintenance tasks.</li> </ul>	<ul style="list-style-type: none"> <li>Foundation models may not always provide the same <b>level of expertise as specialized models</b>, particularly for highly technical tasks such as in-depth analysis of grid infrastructure. In such cases, additional fine-tuning or supplementation with expert systems may be necessary.</li> <li>The broad data used to train foundation models can <b>include biases or outdated information</b>. Therefore, relying solely on a foundation model may result in decisions based on incomplete or skewed information, which highlights the need for thorough validation.</li> </ul>



## Concepts in Large Language Models

This section introduces **pivotal concepts** that enable **LLMs** to traverse a spectrum of applications with precision and adaptability. Each concept is explained on both a high-level intuition and an in-depth technical exposition. Additionally, for each concept, we discuss its main advantages and disadvantages, complemented by three specific use cases to illustrate its practical relevance:

1. **Pretraining** establishes the initial knowledgebase for LLMs, leveraging the transformer architecture to learn a broad linguistic understanding from extensive, unlabeled text datasets.
2. **Fine-tuning** further refines these models, aligning them with specific domain requirements by introducing targeted datasets.
3. **Reinforcement Learning from Human Feedback (RLHF)** marks a paradigm where models iteratively adapt based on human evaluative feedback, aligning their outputs more closely with human expectations.
4. **Connecting LLMs to tools** expands model capabilities, allowing them to access and interact with external tools and computational services and thereby enriching LLM responses with external operational functionalities.
5. **LLM Agents** envision LLMs as proactive entities capable of executing complex tasks through interaction with various tool extensions.
6. **LLM Information Retrieval** combines the generative strengths of LLMs with advanced information retrieval techniques, that significantly enhance the factual accuracy by dynamically incorporating external data sources.
7. **Prompt Engineering** articulates the art of query (prompt) optimization to get desired responses from LLMs.

### Pre-training

Most LLMs at the state of writing are based on the so-called **transformer architecture** [24]. This architecture was originally designed for converting one sequence of tokens into another, e.g., translating texts from one language into another. An important follow-up breakthrough were **generative pre-trained transformers** (GPT [25]), which were simply trained on predicting the next token in a given sequence of tokens (i.e., words, parts of words or characters).

Pre-training gives language models an overall **text understanding** by training it on arbitrary, unlabeled text sequences. Typically, this is achieved through an extremely large corpus of heterogeneous texts. After pre-training, language models can **generate grammatically and semantically meaningful texts**. They also **gather knowledge** from the training data and are able, to a certain degree, to answer questions. Although originally designed for language translation, over the years, it has become more and more evident that, through pre-training, transformers can be applied to a variety of problem-solving tasks they are not explicitly trained on, such as **question answering** or **language inference** and **reasoning** [25, 26, 27, 28].



### Pre-training use cases in the energy sector

#### Foundation Model for subsequent energy-specific tasks

- Create an own foundational language model in which the usage of specific language is enforced (or ruled out) through the selected training documents.
- The model can then be fine-tuned to more specific use cases in subsequent steps.
- E.g., rule out toxic behavior since open language models could be trained based on all kinds of sources.

#### Notice:

- Training a state-of-the-art LLM can be very hardware-intensive and expensive.
- Typically, pre-trained open language models are used instead and fine-tuned to specific tasks.
- Pre-training from scratch can be useful for very specific use cases when smaller models are used.

Potential	Risk
 low potential                      high potential	 low risk                      high risk
<ul style="list-style-type: none"> <li>Training data is not required to be labeled, i.e., it can be comprised of arbitrary text corpora, which are widely available.</li> <li>Allow models to gain an overall knowledge and understanding of language.</li> <li>Rule out unwanted language through the selection of training data.</li> </ul>	<ul style="list-style-type: none"> <li>Pretraining a large state-of-the-art language model is extremely expensive. Hence, typically, an open, pretrained model is used instead and fine-tuned to another domain (“transfer learning”).</li> <li>Requires very large amounts of training data. It can be difficult to ensure quality of the training data and rule out the model learning undesired behavior/text outputs.</li> </ul>

**Fine-Tuning**

Fine-tuning in the context of LLMs represents a crucial stage where pre-trained models are further **refined with specific data** samples to improve their performance on tasks closely aligned with an organization’s unique requirements [22].



**Technically**, fine-tuning involves the continued training of a model that has already been generalized on a vast corpus of data. During the subsequent training phase, the model adapts its pre-learned representations to nuances and intricacies unique to the target domain by leveraging a smaller, domain-specific dataset. Fine-tuning involves striking a careful balance between learning new patterns specific to the dataset and retaining valuable knowledge that has already been learned. This can be particularly challenging when dealing with complex models such as foundation models [29].

Nevertheless, the fine-tuning process comes with **technical complexities and accessibility challenges**. For instance,

models like GPT-4 have restrictive access protocols, which only allow users to communicate via API endpoints [22]. Moreover, the datasets used to train these models are undisclosed, which limits research and development within the AI community and hinders the ability to independently train and refine foundational models. Furthermore, adapting foundation models requires modifying model gradients, which is a more complex process than the simple prompt specification used in in-context learning [30].

In the **energy sector**, this translates into tailoring models to comprehend and accurately use sector-specific terminology and technical formulations. For companies operating in jargon-rich industries, fine-tuning offers a chance to infuse LLMs with the nuanced language of energy systems, regulatory frameworks, and technical specifications unique to their operations. The next subsection will discuss **Reinforcement Learning from Human Feedback** as an alternative to traditional fine-tuning methodologies for training sophisticated LLMs.

Fine-Tuning use cases in the energy sector
<p><b>Technical requirements extraction</b></p> <ul style="list-style-type: none"> <li>Fine-tune LLMs to analyze technical documents and reports.</li> <li>Extract and summarize key technical requirements for grid maintenance and development projects.</li> <li>Streamline project planning and ensure compliance with technical specifications.</li> </ul>
<p><b>Regulatory compliance monitoring</b></p> <ul style="list-style-type: none"> <li>Fine-tune LLMs to understand and interpret the specific language and implications of energy regulations.</li> <li>To scan internal documents and correspondences to ensure they adhere to current laws and guidelines.</li> </ul>

Incident response training simulations	
<ul style="list-style-type: none"> <li>Refine an LLM to create practical training simulations for incident response and provide company-specific feedback to employees.</li> <li>Including potential scenarios such as outages, natural disasters, and other grid emergencies.</li> <li>Compare human responses to a collection of well-handled incidents and provide individual feedback to employees on how to better align with company communication guidelines.</li> <li>LLM can be connected to data sources such as concrete protocols that can be referenced via <i>Information Retrieval</i>.</li> </ul>	
Potential	Risk
 <p>low potential                      high potential</p>	 <p>low risk                              high risk</p>
<ul style="list-style-type: none"> <li>Fine-tuning enables LLMs to become <b>highly specialized in understanding specific documentation</b> and terminology, resulting in more accurate extraction and summarization of technical requirements from reports.</li> <li>An LLM that is fine-tuned with data can <b>more effectively monitor</b> regulatory compliance by comprehending the nuanced language of energy regulations and internal documents.</li> </ul>	<ul style="list-style-type: none"> <li>The process of fine-tuning an LLM requires <b>significant computational resources</b> and expertise, which may entail a substantial investment in both hardware and skilled personnel to manage the fine-tuning process.</li> <li>There is a risk of <b>overfitting</b> the model to specific data, which could result in the model performing well on company data but losing its ability to generalize to</li> </ul>



**Reinforcement Learning from Human Feedback**

**Reinforcement Learning from Human Feedback** (RLHF [31]) improves LLMs on following human instructions and human expectations on generated answers on writing style and overall behavior. Specifically, LLMs are expected to be truthful, to give no toxic or hallucinated answers, and to give answers in a style that is customer-assistant appropriate. The overall goal of RLHF is called *alignment*.

In InstructGPT [31], to perform alignment through RLHF, humans rank different generated answers to the same prompt from best to worst. Based on the given feedback, a “reward model” is trained, which can evaluate what kinds of answers are preferred by users. Finally, the LLM is trained on new prompts to optimize its policy of generating answers to maximize the expected reward. This optimization method is called

“proximal policy optimization” (PPO). Through this overall process, the LLM is aligned to follow the expected behavior, i.e., generating answers like the top-ranked ones.

RLHF through PPO is the basis for conversational chatbots such as ChatGPT, which is based on InstructGPT. A more recently proposed alternative, which is used for the alignment task instead of RLHF, is using **direct preference optimization** (DPO [32]). One major advantage of DPO is that it is a significantly less complex process and much easier to implement in practice while performing similarly or better.

Reinforcement Learning from Human Feedback use cases in the energy sector	
<p><b>Fine-tune to corporate writing style</b></p> <ul style="list-style-type: none"> <li>Employees give feedback for generated texts by ranking different outputs.</li> <li>LM adapts to desired behavior and writing style.</li> </ul>	
<p><b>Adapt agent model to the desired behavior</b></p> <ul style="list-style-type: none"> <li>Develop Agent-based internal applications, e.g., for process management.</li> <li>Adapt the agents to follow desired outputs.</li> </ul>	
<p><b>Question answering for corporate documents</b></p> <ul style="list-style-type: none"> <li>Develop an assistant application that answers questions regarding corporate documents.</li> <li>Adapt the assistant to the desired behavior.</li> </ul>	
Potential	Risk
 <p>low potential <span style="margin-left: 150px;">high potential</span></p>	 <p>low risk <span style="margin-left: 100px;">high risk</span></p>
<ul style="list-style-type: none"> <li><b>Allows models to better follow instructions</b>, i.e., especially useful to fine-tune chatbots, which should follow human instructions, e.g., answering their questions.</li> <li><b>Reduce toxic behavior and hallucinations</b></li> </ul>	<ul style="list-style-type: none"> <li>Requires human feedback from domain experts (depending on the specific use case), i.e., <b>manual data annotation effort</b></li> <li><b>Difficult to implement.</b> Achieving stable results involves careful fine-tuning of hyperparameters in the training process.</li> <li>Consider using DPO instead.</li> </ul>

### Connect Large Language Models to Tools

The integration of LLMs such as GPT-4 with **external tools** is particularly useful for augmenting their utility within specialized sectors. This integration enables LLMs to interact with and leverage diverse databases and software services, expanding their functionality beyond native language processing (NLP) capabilities, as shown in *Figure 7*. From a technical perspective, the challenge is to orchestrate LLMs and interfaces (such as APIs and databases) dynamically. The objective is to improve the model’s output by integrating external, real-time data or performing specific tasks that the LLM cannot perform natively.

Among other resources, *Langchain* provides various out-of-the-box tools that can be connected to ChatGPT APIs [34]. Tools provided allow to screen research literature via arXiv, search local files such as PDFs or Word documents, trigger AWS workflows, or reach out to humans in cases of uncertainty.

The **arXiv paper database** tool for LLMs enables the automated retrieval of scholarly articles. Technically, this integration allows an LLM to execute structured queries against arXiv’s extensive repository, utilizing its API to fetch detailed metadata. This means that upon receiving a query regarding, for example, „grid stability research,“ the LLM can instantaneously procure a list of pertinent publications, complete with their publishing dates, titles, authors, and summaries. This tool allows for swift access to the latest research insights which are instrumental in guiding strategic decisions and maintaining a competitive edge in the energy sector.

Integrating **LLMs with local file system** access tools equips the models with the ability to traverse and analyze internal document repositories (see subsection LLM Information Retrieval for more information). From a technical standpoint, this integration leverages the LLM’s natural language processing strength to parse and understand the contents of various



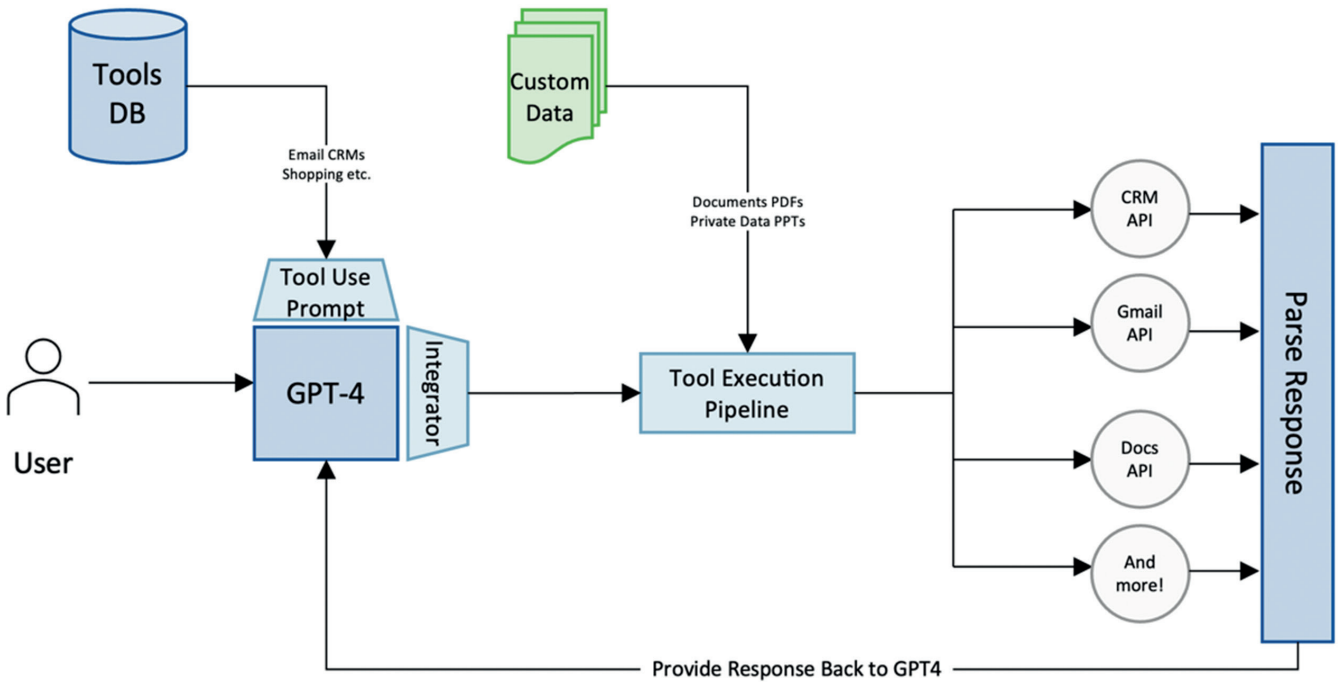




Figure 7: Information flow when connecting LLM to tools, based on [33]

document types, enabling a search capability that is contextual by files. The resultant application is a robust internal knowledge base, where accumulated resources, such as reports, manuals, and policy documents, become readily searchable, streamlining information retrieval, and facilitating knowledge dissemination within the organization.

**Amazon AWS Lambda’s** serverless computing framework can also be combined with LLMs as a tool. This allows the user to invoke serverless functions, such as AWS SageMaker, to perform computation-intensive tasks triggered by the LLM. This way, non-technical employees can trigger Lambda functions, to e.g., interface with real-time grid monitoring services, process the data, and relay updated metrics directly back to the user. This tool connection can form an easy-to-use interface into complex workflows and data science pipelines implemented by technical experts.

Another tool extension that can be plugged into the OpenAI API via *Langchain* is the feedback loop back to the **human**. The tool evaluates the model’s certainty and, in case of doubt, requests more information from the user. This dynamic allows for information to be fed sequentially, especially in complex tasks, instead of collecting all relevant information beforehand.

Recent literature suggests that LLMs can automatically identify the appropriate tool to trigger, such as *GPT4Tools* and *Toolformer* for **self-instructed tool use** [35] [36]. Within specific modalities such as computer vision specific models are already capable of iterative reasoning and connecting multiple models such as Visual Transformers and Diffusion Models in a multi-step process [37]. First, frameworks go beyond by addressing task-agnostic and modality-agnostic task comprehensiveness, such as the OFA framework [38].

Large Language Model Tool use cases in the energy sector	
<p><b>Stay up to date with research</b></p> <ul style="list-style-type: none"> <li>When a user queries about a specific topic like „research on grid stability,“ the LLM can run a query using the <i>arXiv tool</i> extension.</li> <li>Return articles with their publishing date, title, authors, and summaries.</li> </ul>	
<p><b>Local file system search</b></p> <ul style="list-style-type: none"> <li>Navigate the company’s internal repositories to find documents like reports, operational manuals, and policy documents.</li> <li>Form company-specific internal knowledge base and interface.</li> </ul>	
<p><b>AWS lambda for dynamic content</b></p> <ul style="list-style-type: none"> <li>An employee requests data analysis on grid load.</li> <li>Lambda functions trigger AWS sage maker interfacing with real-time grid monitoring services.</li> <li>Return the latest figures and insights to the ChatBot and to the user.</li> </ul>	
Potential	Risk
 <p>low potential <span style="margin-left: 150px;">high potential</span></p>	 <p>low risk <span style="margin-left: 100px;">high risk</span></p>
<ul style="list-style-type: none"> <li>Connecting LLMs to tools like arXiv or AWS Lambda allows the LLM to provide <b>dynamic and contextually relevant responses</b>. They can trigger external Tools and 3rd party applications automatically.</li> <li>LLMs serve as a unified interface for a diverse array of tools and applications, offering a standardized point of interaction. This not only <b>empowers employees to engage with a wider variety of tasks</b> but also enhances their efficiency in individual tasks by streamlining their access to information and resources.</li> </ul>	<ul style="list-style-type: none"> <li>The integration of various third-party tools and APIs with an LLM <b>increases system complexity</b>. This can lead to higher maintenance requirements and the need for specialized staff to manage and troubleshoot the integrations.</li> <li>The performance and reliability of the LLM’s outputs are partly <b>dependent on the third-party tools</b> it is connected to. If these tools experience downtime or disruptions, it can directly impact the functionality of the LLM</li> </ul>

**Large Language Model Agents**

The combination of LLMs with tools allows to regard language models such as chatbots as “agents” which are embedded into a specific environment in which they can perform various actions, communicate with other agents, or make observations. **Agent models** have been widely used in artificial intelligence before, e.g., reinforcement learning, but also other subfields. The potentials of LLMs have helped re-establish interest in concepts from these lines of research, such as **planning and reasoning**.

The overall principle of LLM agents is to present a **task** to the agent via text prompts. **Actions**, such as answering a question,



processing intermediate inputs, or using various tools, are then performed based on LLM outputs and **observations**. Tasks may be broken down into different steps through carefully engineered prompts, together with outputs from the used tools and the intermediate outputs from the LLM itself. Depending on the allowed tools, agents may be able to answer questions, generate, revise, download and execute program code or other machine learning models [39], lookup facts from search engines, databases or knowledge bases, or acquire data required for a given task. Many agent models can perform reasoning based on predefined prompts, reasoning patterns and made observations [40], [41], [42] . Instead of using a single agent, one may incorporate several agents, which are each responsible for different subtasks and which interact with each other.

**Combining several agents** was found to produce more reliable overall results [43].

LLM-based agent models show great potential for automating complex tasks. However, they are still in a state where more research is required. Specifically, being unpredictable in their outputs, agents may be unreliable in consistently solving the desired tasks. One way to mitigate this is to incorporate a

human-in-the loop which supervises the agent’s actions and leads him onto the correct path [43].

The automation of these complex tasks is achieved via a concept called **chains**. They employ a series of functions to architect specialized workflows tailored for a range of applications, notably including entity extraction, document tagging, and enhanced question-answering systems that incorporate citation mechanisms.

Large Language Model Agent use cases in the energy sector	
<p><b>Assistant for data analysis</b></p> <ul style="list-style-type: none"> <li>LLM Agents can help with data pre-processing and plotting.</li> <li>Agents generate code (e.g., for matplotlib).</li> <li>Retrieve required data from external sources through tools.</li> </ul>	
<p><b>Agents as an interactive knowledge base</b></p> <ul style="list-style-type: none"> <li>Different Agents for different business areas.</li> <li>Agents fetch and rephrase relevant info for question answering, e.g., from internal Wiki, SharePoint, etc.</li> <li>Can be more reliable than a single LLM.</li> </ul>	
<p><b>Agents as “Copilots”</b></p> <ul style="list-style-type: none"> <li>E.g., automated Mail Answering.</li> <li>Adding relevant Info/Files to mail from different sources.</li> <li>The user only has to confirm mail.</li> </ul>	
Potential	Risk
 <p>low potential                      high potential</p>	 <p>low risk                      high risk</p>
<ul style="list-style-type: none"> <li>Agents allow for <b>solving tasks through text prompts</b> (e.g., finding and summing up some information from different data sources, drafting a mail to someone, generating, and running program code etc.). A combination of agent models can sometimes <b>solve tasks more reliably</b> than a single LLM.</li> <li><b>Most natural way to interface</b> between human and computer. Potential to replace computer programming and perform arbitrary task solving to a certain degree</li> </ul>	<ul style="list-style-type: none"> <li>Prompts <b>may be misinterpreted</b> by LLMs or leave some ambiguity of the task to solve. LLM-Agents may struggle to solve a task as intended, require several attempts, or not solve at all. Hence, some human supervision and tuning is required.</li> <li>Letting agents run arbitrary code may pose a <b>security risk</b>. Hence, typically, the actions of an agent would be restricted to certain tools and important decisions supervised by a human (“human-in-the-loop”).</li> <li>Risks strongly depend on the use case.</li> </ul>

### Large Language Model Information Retrieval

Retrieval-Augmented Generation (RAG) was first introduced in a 2021 paper by Facebook Meta. It is a groundbreaking concept that surpasses the factual knowledge embedded wit-

hin the parameters of large pre-trained models [44]. The RAG framework synergizes the generative capabilities of LLMs with the precision of information retrieval, setting a new benchmark for state-of-the-art results in a multitude of NLP applications.

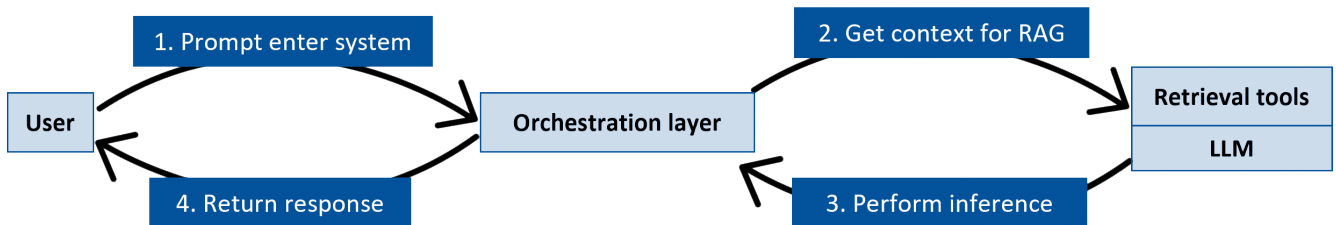


Figure 8: Information flow during Information Retrieval in Large Language Models, based on [45]



The main advantage of RAG systems is their ability to address two major limitations that are inherent in LLM applications: the tendency to generate responses that are factually inaccurate, known as ‘hallucinations’, and the degradation of the model’s relevance over time due to outdated training corpora.

RAG addresses these limitations by dynamically incorporating the retrieval of contemporary external data within the LLM’s generative process, as shown in Figure 8. This integration facilitates access to up-to-date and verifiable external knowledge sources. RAG is especially beneficial for tasks that are complex

and knowledge-intensive. Taking concrete external information into account increases consistency and improves reliability.

The practical implementation of RAG is facilitated through tools such as document loaders, which support a **variety of document types, sources, and formats** [46]. These loaders can process heterogeneous data formats, ranging from structured CSV files to complex PDF documents and extending to communicative data channels such as Slack. An LLM equipped with this technology can systematically navigate and extract information from various storage solutions, including S3 buckets, publicly accessible web domains, and other databases.

Information Retrieval use cases in the energy sector
<p><b>Company internal expert finder</b></p> <ul style="list-style-type: none"> <li>■ Retrieve information on previously executed projects with similar scopes.</li> <li>■ Identify colleagues working on similar topics and form synergies.</li> </ul>
<p><b>Optimize organization functions and structure</b></p> <ul style="list-style-type: none"> <li>■ Connect various function and organizational-related working folders.</li> <li>■ Identify overlapping efforts or redundancies by comparing project embeddings.</li> </ul>
<p><b>Summarize and reason about project pain points</b></p> <ul style="list-style-type: none"> <li>■ Connect environmental reports, geographical data, project drafts, and reasons about cause and relation via knowledge graphs.</li> <li>■ Highlight potential risks and bottlenecks with direct reference to the sources.</li> </ul>

Potential	Risk
 <p data-bbox="236 416 373 448">low potential</p> <p data-bbox="552 416 699 448">high potential</p>	 <p data-bbox="911 416 994 448">low risk</p> <p data-bbox="1171 416 1254 448">high risk</p>
<ul style="list-style-type: none"> <li>■ RAG can enhance the <b>factual consistency</b> of the language model’s outputs. It adds a layer of verification to the generative process.</li> <li>■ Since RAG uses real-time data retrieval, it <b>reduces the impact of out-of-date</b> training data that can affect the performance of traditional language models. Also, it can answer beyond what was seen during training.</li> </ul>	<ul style="list-style-type: none"> <li>■ As the SharePoint grows, the volume of data to be processed will increase, potentially impacting the performance and scalability of the RAG system, which is limited by the context window of a currently deployed model.</li> <li>■ If the RAG system is overly relied upon without proper human oversight, there could be a risk of missing nuanced context or emergent issues not captured by the existing data or missed by the context window.</li> </ul>



### Large Language Model Prompt Engineering

As stated in the Google developers guide [47], “**Prompt engineering** is the art of asking the right question to get the best output from an LLM. It enables direct interaction with the LLM using only plain language prompts.” Prompt engineering requires a nuanced understanding of both the model’s capabilities and the intricacies of natural language, enabling effective communication with the LLM without complex programming constructs. Concrete concepts such as in-context learning and chain-of-thoughts prompting can help to advance prompts.

**In-context learning**, coined by the Center for Research on Foundation Models, suggests that an LLM can be customized for specific downstream tasks by providing a prompt - a brief natural language instruction that encapsulates the task [22]. By providing concrete examples on which output the LLM should provide on a given input, a prompt where the task is defined (considered zero-shot) can be extended to form a One-, few-, and multi-shot depending on how many examples are provided. In contrast to LLM fine-tuning (see subsection Fine-Tuning) in-context learning is less resource-intensive and present a viable alternative, offering a balanced tradeoff between accuracy and computational efficiency. Most recent advancements in research extended prompting towards Multimodal Reasoning and Action, as shown in the framework MM-REACT [38].

**Chain-of-thought prompting** is a more nuanced form of prompt engineering that has been highlighted for its potential to enhance the reasoning abilities of LLMs [41]. This method involves asking the LLM to provide its thought sequence leading up to a final answer. The model guides itself through the cognitive steps required to resolve complex, multi-stage problems. The communicated intermediate steps also allow the user to question and validate the process that led to the final answer. An extensive overview on the interplay of fine-tuning LLMs on chain-of-thought prompting and Instructions with examples is provided in Google’s paper “*Scaling Instruction-Finetuned Language Models*” [23].

The practical implementation of prompt engineering is supported by libraries that offer to pre-define **prompt templates** [48]. Such tools permit the specification of variable inputs within expertly engineered prompts. In addition, LangChain enriches this ecosystem with its capacity to manage diverse message types—ranging from AI-generated messages to human interactions—allowing for the customization of prompts in accordance with the communicative context and the intended recipient’s role. This flexibility empowers users to instruct LLMs in a manner that is both contextually appropriate and technically precise, optimizing the interaction for clarity and efficacy. All while only needing to keep a minimal, task-specific prompt for each distinct task, which, when applied, enables mixed-task inference while utilizing the same pre-trained model [30].

Prompt Engineering use cases in the energy sector	
<p><b>Prompt templates for routine queries</b></p> <ul style="list-style-type: none"> <li>■ Experts engineer and define the prompt for collecting, summarizing, and presenting information.</li> <li>■ An individual request is filled by an input variable.</li> </ul>	
<p><b>Learn requirement engineering in-context</b></p> <ul style="list-style-type: none"> <li>■ Define energy sector-specific technical requirement information.</li> <li>■ Provide in-context examples of mappings.</li> <li>■ Screen tenders and generate energy sector-specific report.</li> </ul>	
<p><b>Chain-of-thought templates to cover various facets</b></p> <ul style="list-style-type: none"> <li>■ Define areas that need to be covered when working on a recurring complex problem.</li> <li>■ Have an LLM challenge all facets by an engineered prompt.</li> </ul>	
Potential	Risk
 <p>low potential                      high potential</p>	 <p>low risk                      high risk</p>
<ul style="list-style-type: none"> <li>■ Prompt engineering allows for the crafting of precise queries that can elicit specific information and actions from an LLM. A <b>broad group of employees can be empowered</b> to benefit from LLMs.</li> <li>■ By prompting methods, one can achieve a favorable <b>balance between accuracy and efficiency</b> without the extensive resources required for fine-tuning.</li> </ul>	<ul style="list-style-type: none"> <li>■ Developing effective prompts is an art that <b>requires a deep understanding of LLMs</b> and how they interpret and generate language. One will need experts who can engineer prompts that lead to the desired outcomes.</li> <li>■ While chain-of-thought prompting can guide an LLM through multi-step problems, it <b>may not always capture the full complexity of certain tasks</b>, such as those involving high-stakes decision-making or intricate technical analyses that one might require.</li> </ul>



## GenAI-Based Use Cases for TenneT

This section centers on TenneT and discusses specific GenAI-based use cases for TenneT in more detail. Collaboratively developed with TenneT employees in a workshop, the following four use cases progress through a structured methodology. Initially, based on a current problem at TenneT, the corresponding users and the call for action are identified. Subsequently, requirements and their coverage are developed, culminating in proposed solutions.

To facilitate systematic evaluation and comparison, evaluation criteria are jointly formulated with the TenneT employees. Subsequent analysis involves assessing the potential and associated risks of each use case. The following criteria are evaluated:

**Expected value** serves as a multi-faceted criterion encompassing both tangible and intangible benefits. It encapsulates the anticipated contributions of GenAI to TenneT's strategic imperatives, considering not only the direct financial impact but also the potential to enhance the company's competitive positioning within the industry. This criterion also summarizes the added value of GenAI from elevating public awareness of TenneT's initiatives to the strategic relief of critical resources. The scalability of a solution – its ability to expand and adapt to a growing target audience – is also a vital component, ensuring long-term alignment with corporate-level strategies.

The criterion **Effort** encompasses various dimensions of resource investment required for successful implementation and sustained operation. It includes upfront financial investments, which are crucial for procuring the necessary infrastructure and technology to deploy GenAI solutions. Also considered are the costs associated with the initial setup and subsequent maintenance, underscoring the long-term financial commitment that organizations must be prepared to undertake. It reflects the operational expenditure, such as the personnel costs for training and managing the GenAI systems, and the potential need for continuous financial outlays post-implementation to ensure the solutions remain effective and up-to-date. This also ties into the cost of change management and educational initiatives necessary to integrate GenAI into existing workflows, highlighting the importance of considering both immediate and ongoing costs when assessing the feasibility and value of GenAI applications in the energy sector.



The evaluation criterion of the **Availability of Resources** pivots around the accessible human and technical capital necessary to realize GenAI projects within TenneT. It underscores the significance of having skilled personnel, both in-house

and external experts, available to support the integration and maintenance of GenAI technologies. This criterion also factors in the Time to Deployment — the duration it takes from initiating a GenAI project to its full operational status — which is critical for meeting strategic timelines and achieving measurable impact. This factor also considers the necessity of continuous IT support and the compatibility of GenAI with existing IT infrastructure, acknowledging the importance of seamless technological integration. Quantifiable measures discussed are the full-time equivalent (FTE) resources required to ensure the sustained success of GenAI applications, highlighting the importance of dedicated teams for the development, deployment, and evolution of these systems in a manner that aligns with TenneT's operational and strategic imperatives.

The **Compliance** category within the evaluation framework covers several critical aspects that ensure the safe and lawful application of GenAI technologies. This includes adherence to data security and protection standards, aligning with GDPR and cybersecurity requirements to safeguard against data breaches and ensure user privacy. Trust and truthfulness are paramount in establishing the credibility of GenAI applications, ensuring that the technology operates within ethical boundaries and generates reliable outputs. The category also covers the potential governance issues that may arise, emphasizing the importance of GenAI solutions being compliant with both regulatory frameworks and TenneT's internal guidelines.

**Maturity** refers to the level of sophistication and readiness of AI technologies. Questions about where the GenAI solutions are hosted are a critical factor in operational efficiency and scalability. The availability of commercial products plays a crucial role when discussing the accessibility and readiness of AI solutions for immediate use. Expected acceptance considers how likely users are to embrace the new technology, taking into account factors like ease of communication and the complexity of the AI solution. Interoperability with partners stresses the importance of AI systems being able to work seamlessly with different platforms and technologies.

On a more holistic level, TenneT should consider to what extent the **Environmental Implications** of GenAI align with the company's philosophy on energy consumption and CO2 emissions. The incorporation of these models must be weighed against their carbon footprint and energy demands, ensuring they cohere with TenneT's commitment to sustainability and environmental responsibility. This assessment should extend beyond the immediate energy requirements for training and

Use Case 1: Public Affairs & Communication Support	
<b>Public Affairs &amp; Communication Support</b>	
<b>Problem statement</b>	
<p>The integration of LLMs and Diffusion models into the daily operations of the Public Affairs &amp; Communication department could yield significant benefits. Currently, employees are using these technologies independently or in small groups, resulting in a variety of practices without a unified framework. There is a clear opportunity to enhance content creation processes at TenneT by collectively understanding and sharing best practices. This opportunity is currently not being fully realized due to varied individual approaches and a general gap in GenAI literacy.</p>	
<b>Outline solution</b>	
<p>The proposed solution involves implementing prompt engineering tutorials, establishing a prompt Wiki, providing prompt engineering support, and creating a front end with templates and best practices to streamline content generation. Additionally, the exploration of diffusion models for marketing-oriented figure and image generation, pending the resolution of legal inquiries, offers promise.</p>	
<b>Potential</b>	<b>Risk</b>
	
<p>low potential                      high potential</p>	<p>low risk                              high risk</p>

utilizing GenAI models to include the latent environmental impact associated with the supporting infrastructure. Factors such as the operational efficiency of data centers, the energy intensity of cooling systems, and the renewable credentials of the power supply are all pivotal in understanding the full ecological footprint of GenAI implementations.

The main **stakeholders** for this use case are the employees of Public Affairs & Communication (PUC) at TenneT. They are responsible for generating content on a daily basis. However, there is a lack of GenAI literacy within TenneT, resulting in different approaches being taken by different parties when employing prompts for LLMs to address similar tasks. Therefore, PUC’s content generation efforts require a coordinated approach.

In response to this challenge, as a call for action, a multi-faceted approach is warranted. The technical perspective is provided in Figure 9. Firstly, an **educational program** aimed at enhancing GenAI literacy among PUC employees is recom-

mended. This initiative should be complemented by the dissemination of best practices, including the provision of **prompt templates**, to standardize content generation methodologies across TenneT. Prompt engineering tutorials and a prompt Wiki to facilitate skill development and knowledge sharing among PUC employees. Additionally, the implementation of diffusion models for marketing-oriented figure and image generation holds promise.

In terms of **requirements**, content generation has to be both fast and error-free, while adhering to the formal tonality characteristic of TenneT communications. Moreover, the generated content should be perceived as human-created to ensure its resonance with the intended audience. Given the legal considerations inherent in content creation, particularly regarding copyright, personal rights, and GDPR compliance, an in-house solution is indispensable.

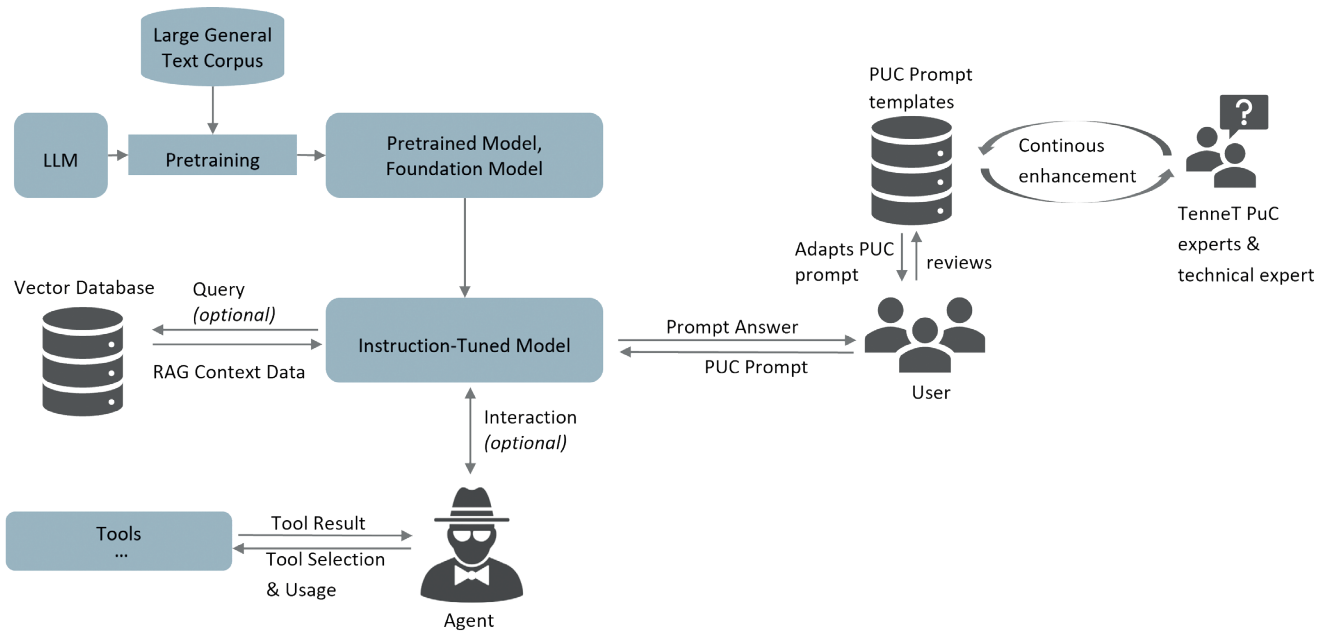




Figure 9: Concept technological solution Use Case 1: Public Affairs & Communication Support

In summary, following this strategy will improve PUC’s ability to generate content, ensuring the timely delivery of high-quality, legally compliant material that resonates with its audience

while maintaining the formal tone synonymous with TenneT communications.

Potential evaluation dimension	Potential
<p><b>Expected value</b> Enhance content generation efficiency and quality; can be adeptly deployed across diverse communication channels, addressing a broad spectrum of target groups, thereby ensuring tailored engagement and enhanced interaction experiences for each audience segment.</p>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>
<p><b>Maturity</b> Prompt engineering tutorials and support, are mature approaches; exploration of diffusion models for marketing-oriented figure/image generation may be in a less mature state, pending resolution of legal considerations.</p>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Risk evaluation dimension	Risk
<p><b>Effort</b> Developing and disseminating educational materials, establishing support systems; potentially navigating legal intricacies surrounding diffusion models.</p>	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<p><b>Availability of resources</b> Pre-trained models are available out of the box; personnel for developing tutorials and support systems; potential funding for legal consultations regarding diffusion models.</p>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<p><b>Data availability</b> Vast historical repository of PUC texts is available for in-context fine-tuning; data control is ensured by hosting instances of open source LLMs on trusted infrastructure.</p>	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<p><b>Compliance</b> Clear guidelines for incorporating target group information; copyright; GDPR.</p>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Use Case 2: Interactive Corporate Knowledge Base	
<b>Interactive Corporate Knowledge Base</b>	
<b>Problem statement</b>	
TenneT employees need an integrated system to efficiently locate and utilize project-relevant information and documents scattered across multiple platforms and formats to quickly become experts in their field of work.	
<b>Outline solution</b>	
An LLM-based system which can compile and reformulate relevant information from multiple data sources (e.g., Documents, SharePoint, OneDrive, etc.). The system has only access to information with an adequate security level to avoid compliance violations. The system uses a RAG-based approach for reliable and consistent answers.	
<b>Potential</b>	<b>Risk</b>
	
low potential                      high potential	low risk                              high risk

The main stakeholders for this use case are TenneT employees, particularly those who are new or transitioning to new projects and topics. These employees face the crucial task of quickly becoming experts in their respective fields. The dispersion of necessary information across multiple platforms, such as SharePoint and OneDrive, and its presentation in varied formats like PPTX and PDF, exacerbates the challenge of finding job-specific information.

Identifying relevant documents is complex and resource-intensive and is further complicated by language and terminology barriers. TenneT, therefore, needed a streamlined approach to onboarding and information retrieval.

An efficient and user-friendly interface to the corporate knowledge base is essential. This interface should enable quick access to necessary information, saving valuable time for employees. Furthermore, project-specific information, along with useful resources, will greatly assist employees in comprehending and excelling in their roles.

Regarding requirements, accessing information from the corporate knowledge base should be easy, despite the scattered

nature and varying formats of the data. The system should also enable interactive dialogues with the knowledge base and allow for the creation of personalized knowledge bases for individual tasks and projects. Open access to information for enriching the corporate knowledge base, robust knowledge management, and compliance assurance are also crucial.

To address these requirements, the proposed solution involves developing a knowledge management system that uses Retrieval Augmented Generation (RAG) to handle corporate, project, and personal knowledge bases effectively. Additionally, a chat system powered by an instruction-tuned LLM, further refined with energy domain-specific content, will be deployed to facilitate user interaction. The integration of tool agents will allow TenneT employees to connect to external data sources and access reliable information efficiently.

This solution equips employees with the necessary tools to become field experts, enhancing productivity and contributing to the company's overall success while maintaining compliance and data integrity.

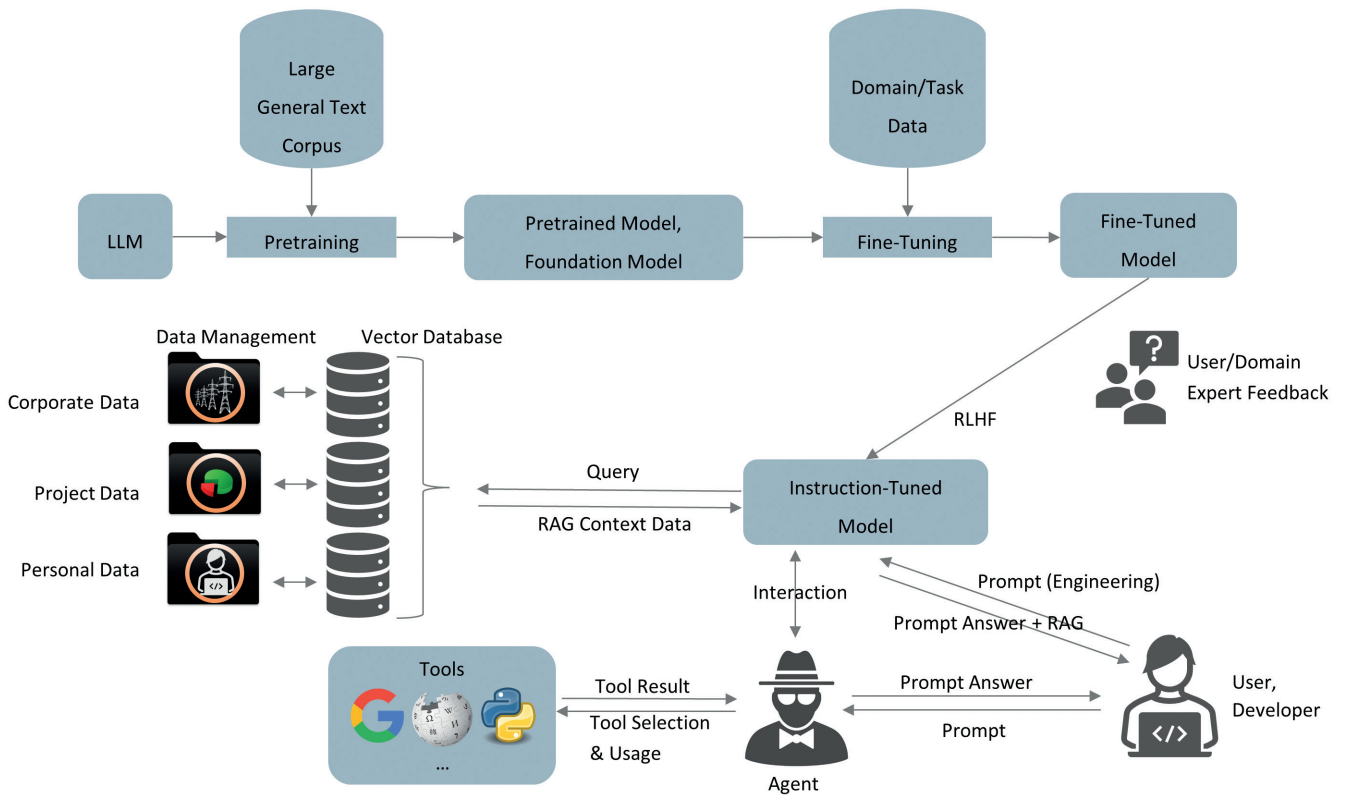


Figure 10: Concept technological solution Use Case 2: Interactive Corporate Knowledge Base

Potential evaluation dimension	Potential
<p><b>Expected value</b> High value is expected. Overall work efficiency is increased, many employees are targeted.</p>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
<p><b>Maturity</b> Vector databases and RAG are well-researched and used in practice. Many different database vendors exist to choose from.</p>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>
Risk evaluation dimension	Risk
<p><b>Effort</b> Connecting various sources, creating a vector database that scales to TenneT company size and data management for project and personal data. The database must be synchronized with new or removed documents. LLM access and overall system maintenance creates running costs. Fine-tuning LLM for domain-specific language and terms.</p>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<p><b>Availability of resources</b> Pre-trained models are available out of the box. Personnel for development of the system. RAG and chat solutions are available.</p>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<p><b>Data availability</b> Files already present on, e.g., SharePoint and OneDrive, data on security levels/access rights can partly be derived from with whom it is shared.</p>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>
<p><b>Compliance</b> For files and information which are already shared, no compliance problems arise. It has to be ensured that project and personal information are shielded by an identity and access management system.</p>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Use Case 3: Energy Sector Professor	
<b>Energy System Professor – explain the energy system</b>	
<b>Problem statement</b>	
Development of a clear concept that explains the impact of the energy transition, especially to users of the energy system and the public in general. The aim is to illustrate and explain how the energy transition affects directly impacted parties. Focus is the expected economic and infrastructural consequences.	
<b>Outline solution</b>	
<p><b>At the beginning:</b> The system generates content based on GenAI, including explanations, graphics, and videos, for communication purposes. It can be used in combination with PUC support. It connects various energy-related data sources.</p> <p><b>In the long term:</b> An interactive chat system to allow those affected to ask questions related to technical and legal regulations. The system will reference documents and provide necessary forms.</p>	
<b>Potential</b>	<b>Risk</b>
<p>low potential                      high potential</p>	<p>low risk                              high risk</p>

The energy sector is undergoing a transformative challenge due to the energy transition. As the shift towards more sustainable energy sources continues, significant changes are taking place within the energy system that affects consumers, producers, and prosumers alike. These changes are complex and have wide-ranging implications, making it difficult for non-experts to fully understand their impact on their business models and daily lives. Therefore, it is essential to provide a clear and accessible explanation of the energy transition and its effects.

To achieve this, a system should be developed that simplifies the complexity and makes the subject matter tangible and understandable. One way to start is by using a content generation system that utilizes GenAI to create explanations, graphics, and videos for communication purposes. The Public User Communication (PUC) framework could utilize this system to integrate various energy-related data sources and provide a comprehensive view.

The long-term goal is to implement an interactive chat system that enables users to inquire about technical and legal aspects of the energy transition. This system would reference relevant documents and facilitate the completion of necessary forms. It would provide a seamless support experience.

The general public, energy consumers, producers, and prosumers can benefit from a personalized explanation and support system. This is especially important given the complexity of the energy system, particularly during the energy transition. To ensure successful adoption of the energy transition, affected users require personalized explanations and support. Key changes, such as those outlined in the Net Development Plan (NEP), must be communicated in an understandable manner to the general public.

To meet these needs, a public information service should be developed that disseminates information to the public and engages in dialogue. An expert assistance system would offer personalized scenarios of changes and implementation suggestions, while a support system would aid with form completion.



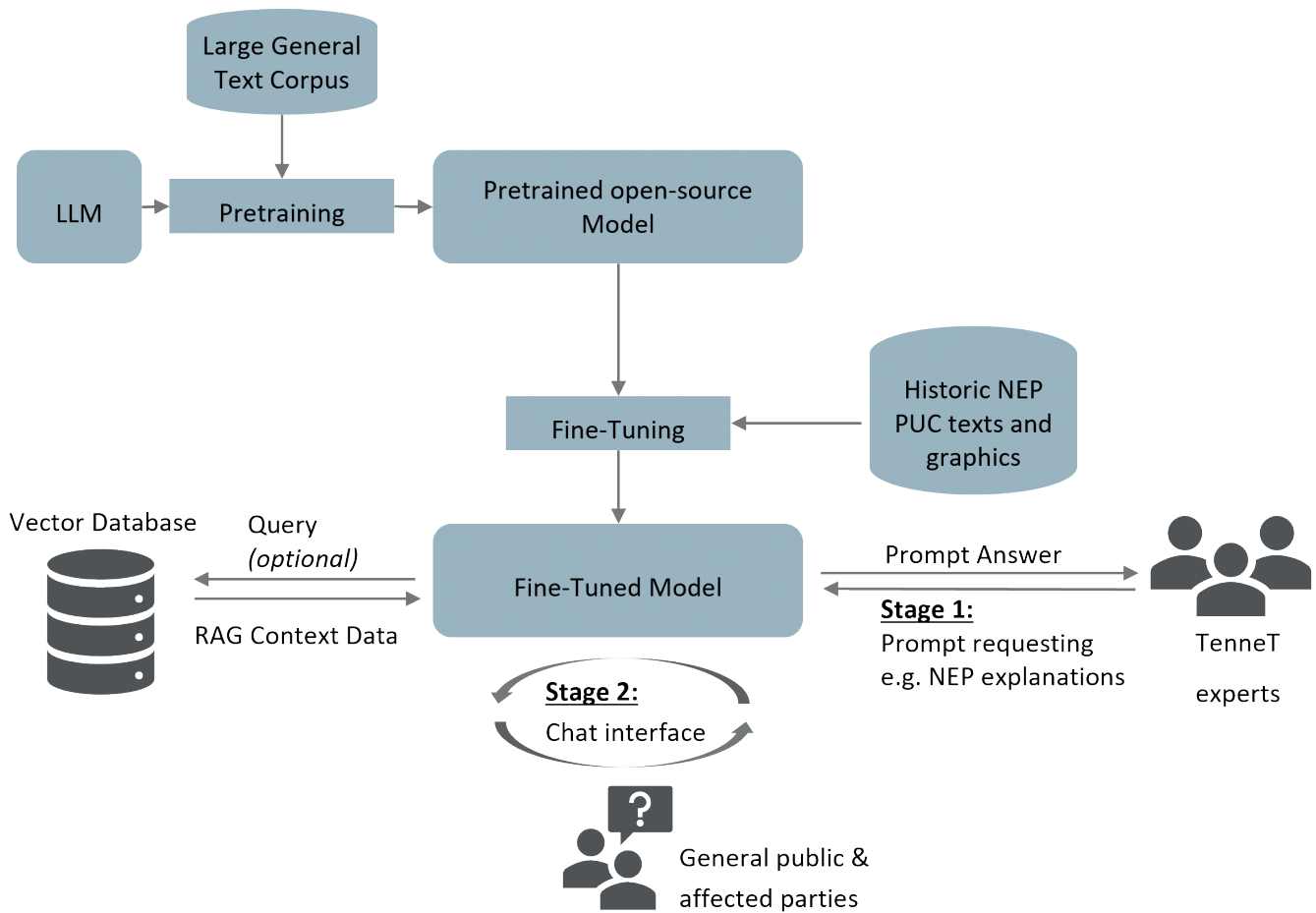


Figure 11: Concept technological solution Use Case 3: Energy Sector Professor

In conclusion, the proposed solution includes a system capable of answering technical, legal, and economic questions within the energy system. It would link information from various sources for personalized understanding and develop task-specific recommendations for the next steps. It is

important to note that the system adapts its responses to the user’s level of experience. This ensures that everyone, regardless of his or her expertise, can navigate the changing landscape of the energy sector.

Potential evaluation dimension	Potential
<b>Expected value</b> Better communication and better support of customers	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>
<b>Maturity</b> Vector databases and RAG are well researched and used in practice. Many different database vendors exist to choose from, multimodal approach is newer but available	
Risk evaluation dimension	Risk
<b>Effort</b> Connecting various sources, creating a vector database which represents economic and infrastructural regulations. The database must be synchronized with changes in regulations. Setup multimodal system and overall system maintenance and running costs. Fine-tuning LLM for explaining energy system topics.	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>

Risk evaluation dimension	Risk
<b>Availability of resources</b> Pre-trained models are available out of the box but have to fine-tuned. Personnel for development of the system, fine-tuning, document curation. RAG and chat solutions are available.	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<b>Data availability</b> Regulation documents are available, standards (i.e. DIN, IEC,..) potentially behind a paywal.	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<b>Compliance</b> External Chat system must comply with data privacy regulations.	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

**Use Case 4: Grid Maintenance Agent**

**Grid Maintenance Agent**

**Problem statement**

The manual inspection of power grid infrastructure, often conducted through labor-intensive methods such as reviewing drone or helicopter videos, presents challenges in detecting problems and anomalies, amplified by the scarcity of reliable training data for automated maintenance support.

**Outline solution**

The solution involves collecting training data, using GANs and diffusion models to augment and refine images, and training classical CNN models for classification. By quantifying uncertainty in the detection process, it facilitates the seamless integration of human experts into the workflow, minimizing the likelihood of misclassification.

Potential	Risk
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
low potential <span style="margin-left: 100px;">high potential</span>	low risk <span style="margin-left: 100px;">high risk</span>

In the field of power grid maintenance, ensuring the reliability and safety of infrastructure is essential. However, the manual inspection of drone or helicopter videos to detect anomalies poses significant challenges at TenneT in terms of resource intensity and cost. Moreover, the lack of robust training data further complicates efforts to implement reliable automated maintenance support. The current situation requires action to simplify maintenance processes, decrease dependence on manual labor, and reduce associated costs. The solution moves

towards automation, using advanced image analysis techniques to improve the efficiency and effectiveness of maintenance operations.

Key requirements for the solution include the ability to estimate uncertainty, detect faults accurately, and provide an interpretable data analysis to the maintenance experts.

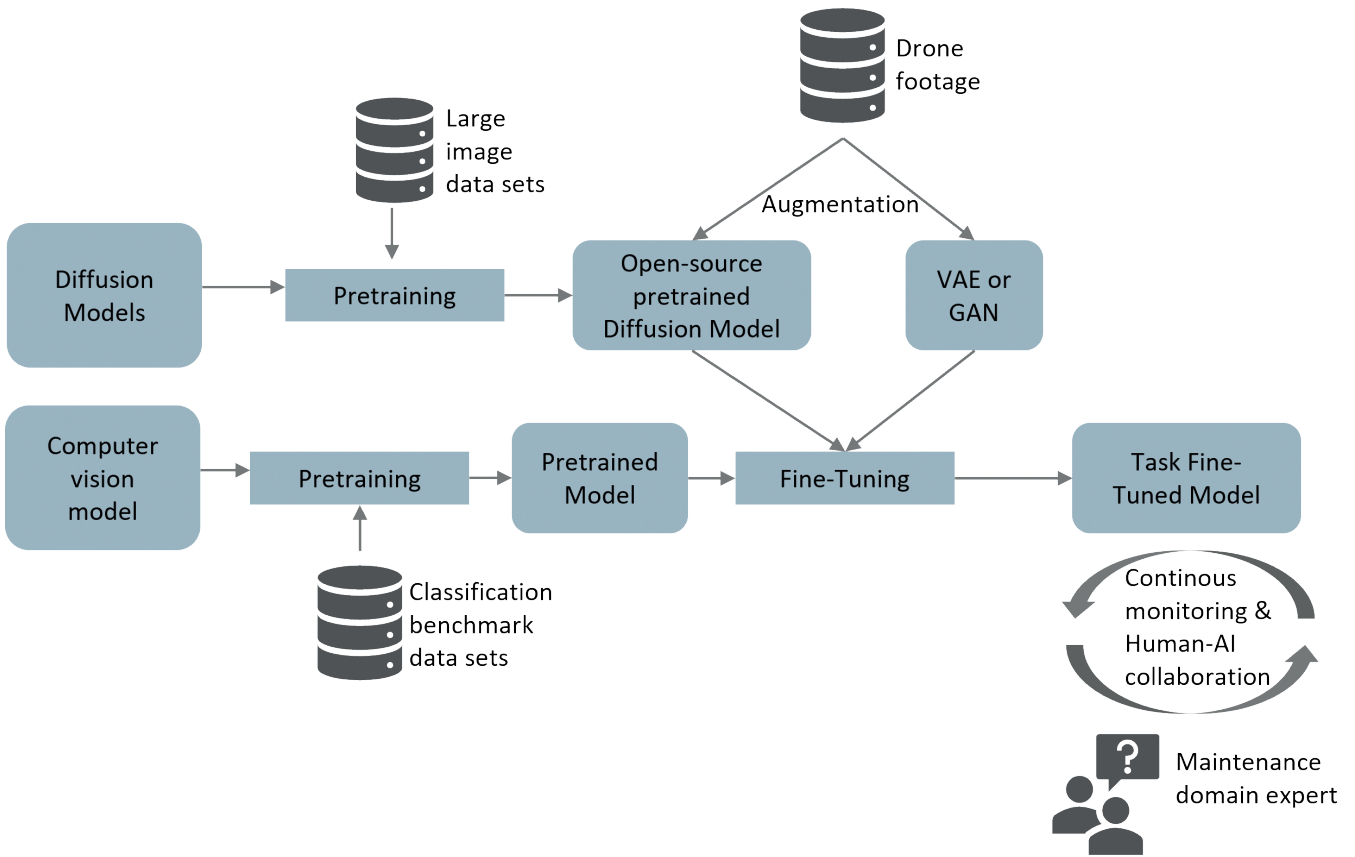


Figure 12: Concept technological solution Use Case 4: Grid Maintenance Agent

To address these requirements comprehensively, the following solution is proposed, shown in *Figure 12*. The initial phase encompasses the acquisition of high-quality training data through recordings obtained via drones or helicopters. Additionally, altering the images and generating additional training footage can enhance classification performance and facilitate more accurate anomaly detection. Besides more traditional GAN architectures, diffusion models can be applied to generate

training images of characteristics that have not been observed before. Furthermore, additional training data will be generated through GANs or diffusion models to augment the dataset and improve model performance. An automated classification pipeline will then screen newly recorded footage and preselect images with high classification uncertainty, signaling the need for expert intervention.

Potential evaluation dimension	Potential
<b>Expected value</b> Cost savings and improved infrastructure reliability.	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>
<b>Maturity</b> On the one hand, well-established model architectures such as CNN and Transformers, but on the other hand, rather new concepts such as diffusion models.	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Risk evaluation dimension	Risk
<b>Effort</b> Data collection, algorithm development, model training, and integrating the solution into existing maintenance workflows.	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Risk evaluation dimension	Risk
<b>Availability of resources</b> Access to necessary resources, skilled personnel, computing infrastructure; higher up-front investment and time to deployment.	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<b>Data availability</b> The availability of high-quality training data is critical for developing accurate and reliable models for automated maintenance support. Hosting the data and trained model instances on trusted infrastructure ensures data control.	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<b>Compliance</b> Recorded video footage and data processing must comply with data privacy regulations.	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

## Conclusion & Recommendations for Action

The proposed use cases in this study, ranging from Public Affairs & Communication Support to explaining the energy system, discuss the potential impact of generative artificial intelligence (GenAI), in more complex facets of TenneT’s operations. They underscore the importance of a knowledgeable

and engaged workforce in realizing the full potential of GenAI. Based on the presented concrete GenAI use cases for TenneT the general recommendation can be summarized as shown in Figure 13. With a temporal resolution, the process begins by enabling employees and progresses towards complex use-case development.

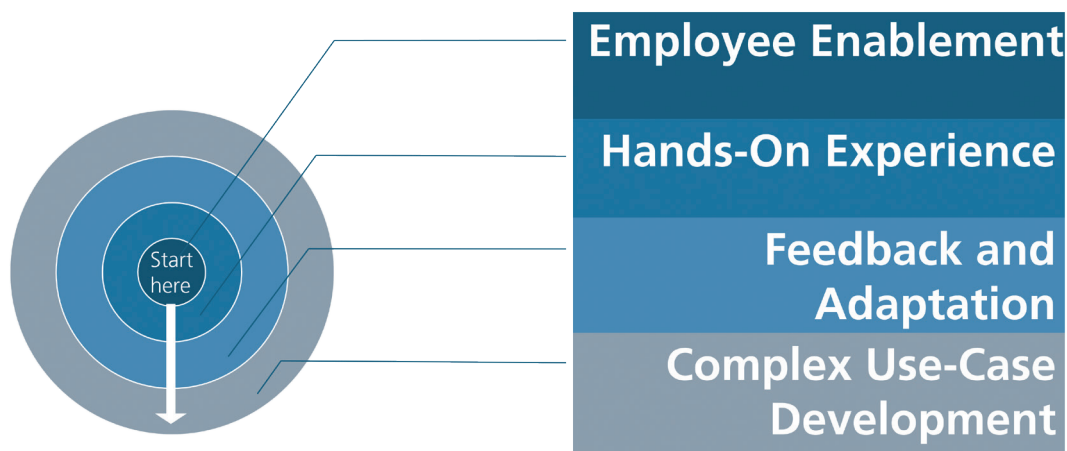


Figure 13: TenneT GenAI paradigms - Cultivating GenAI proficiency by establishing four GenAI paradigms

- 1. Employee Enablement:** Implementing educational programs and practical workshops to build GenAI literacy across the workforce, ensuring employees can effectively utilize Large Language Models (LLMs) and foundation models to enhance productivity.
- 2. Hands-On Experience:** Launching a series of ‘Fireflies’-small-scale, low-effort projects that serve as quick wins to demonstrate the value of GenAI. This approach encourages employees to apply their new learnings in manageable, real-world tasks, fostering confidence and building momentum towards more significant ‘Light-house’ projects. This methodical progression ensures a gradual but tangible enhancement in handling the technology’s practical applications.
- 3. Feedback and Adaptation:** Establishing feedback mechanisms for employees to share insights on GenAI use, which can be leveraged to refine and customize applications to meet TenneT’s operational needs better.
- 4. Complex Use Case Development:** Encouraging employees to participate in the development and advancement of GenAI applications, enabling TenneT to tackle complex challenges with innovative solutions.

In the fast evolving landscape of GenAI, employee enablement at TenneT plays a central role. The immediate adoption of GenAI for low-hanging fruits in everyday tasks hinges on the extensive potential of foundation models and, more specifically, LLMs. By equipping employees with a fundamental understanding and hands-on experience with these technologies, TenneT can capture immediate operational efficiencies and set a foundation for addressing more intricate use cases.

Empowering TenneT’s workforce with basic GenAI knowledge is not just about improving operational capabilities; it’s also about fostering a culture of acceptance and innovation. A workforce adept in GenAI can leverage the technology’s expansive capabilities, from enhancing routine tasks to contributing to more elaborate solutions. Furthermore, employee fluency in GenAI will lead to more nuanced feedback, which is essential for the tailored adaptation of these technologies within TenneT’s unique operational context.

In addition to establishing ideal conditions, it is vital to have a suitable roadmap for introducing GenAI technology in a way that enables TenneT to learn and evolve with it. It is advisable to gradually increase the complexity of the realized application and approach the core tasks of TenneT as a grid operator in a stepwise manner. The developed use cases can be used to realize this. Therefore, we suggest the following roadmap, shown in Figure:

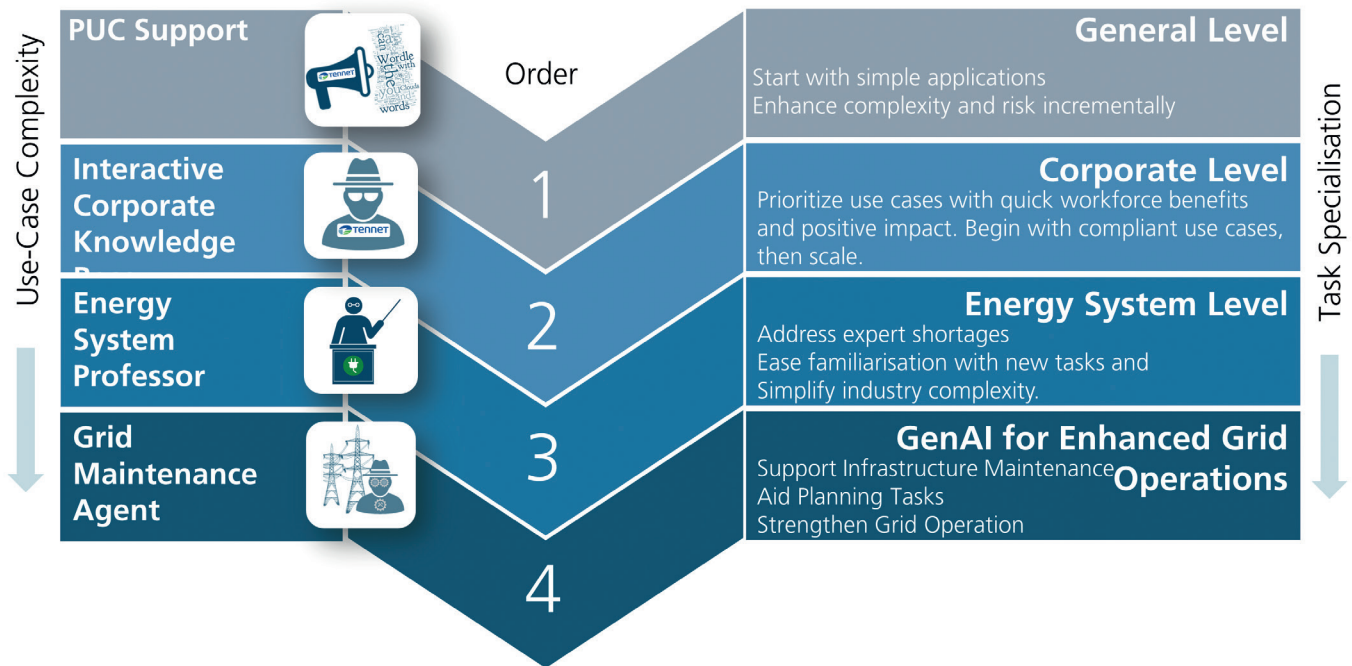


Figure 14: Roadmap for Implementing Generative AI Use-Cases

**1. General Level: Incremental Complexity and Risk Management**

The first step is to introduce GenAI through simple applications that have clear benefits and low risk which applies to everyone, independent of businesses. This gradual approach allows for a controlled environment where the technology's impact can be assessed and managed.

**Use case:** PUC Support

The use case involves utilizing tutorials, prompt Wiki, and a front-end with templates and best practices to streamline content generation. The implementation is straightforward and can quickly benefit the PUC team with minimal risk of errors, as the results are chosen and utilized by PUC experts.

**2. Corporate Level: Workforce Benefits and Compliance**

The next step is to integrate GenAI within corporate structures at scale to streamline processes for large groups of employees. The focus is on identifying use cases that offer immediate benefits to the workforce while ensuring compliance with regulatory standards.

**Use case:** Interactive Corporate Knowledge Base

The use case employs a data management system for corporate, project, and personal knowledge, as well as a chat system for conversing about the knowledge. The availability of corporate knowledge is greatly improved, and compliance is easily managed, making a significant impact on the workforce.

**3. Energy System Level: Addressing Expert Shortages**

At this stage, industry-specific tasks should be addressed by leveraging GenAI, for example, to mitigate expert shortages, help employees become familiar with new tasks in the changing energy system, and simplify the industry's inherent complexity.

**Use case:** Energy System Professor

The use case employs GenAI to clarify the impact of the energy transition for users of the energy system and the general public. It simplifies the complexity of the energy system for its users.

**4. Grid Operation Level: Enhanced Grid Operations**

Finally, as GenAI matures and gains the trust of the organization, it will be integrated into more critical operations. These operations should be prioritized based on their criticality, starting with infrastructure maintenance, followed by grid planning and grid operation strengthening.

**Use case:** Grid Maintenance Agent

GenAI is used to generate training data for assisted inspections of power grid infrastructure via drone or helicopter videos. It efficiently supports experts in detecting anomalies during maintenance of the grid infrastructure.

Through this progressive roadmap, TenneT can effectively integrate GenAI into its operations. The integration will start with simple, low-risk tasks and progress to more complex and impactful applications. This approach ensures a smoother transition and maximizes the technology's potential.



## References

- [1] "The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani)." Accessed: Feb. 15, 2024. [Online]. Available: <https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html>
- [2] "The economic potential of generative AI: The next productivity frontier." Accessed: Feb. 15, 2024. [Online]. Available: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#/>
- [3] I. Goodfellow et al., "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2014. Accessed: Apr. 05, 2023. [Online]. Available: [https://papers.nips.cc/paper\\_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html](https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html)
- [4] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, Jul. 2017, pp. 214–223. Accessed: Apr. 18, 2023. [Online]. Available: <https://proceedings.mlr.press/v70/arjovsky17a.html>
- [5] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dezember 2017, pp. 5769–5779.
- [6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2242–2251. doi: 10.1109/ICCV.2017.244.
- [7] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets." arXiv, Nov. 06, 2014. doi: 10.48550/arXiv.1411.1784.
- [8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks." arXiv, Nov. 26, 2018. doi: 10.48550/arXiv.1611.07004.
- [9] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks." arXiv, Mar. 29, 2019. doi: 10.48550/arXiv.1812.04948.
- [10] A. Asperti, D. Evangelista, and E. L. Piccolomini, "A survey on Variational Autoencoders from a GreenAI perspective." arXiv, Mar. 01, 2021. doi: 10.48550/arXiv.2103.01071.
- [11] M. Mathieu, J. Zhao, P. Sprechmann, A. Ramesh, and Y. LeCun, "Disentangling factors of variation in deep representations using adversarial training." arXiv, Nov. 10, 2016. doi: 10.48550/arXiv.1611.03383.
- [12] I. Higgins et al., "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," presented at the International Conference on Learning Representations, Nov. 2016. Accessed: Dec. 11, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/beta-VAE%3A-Learning-Basic-Visual-Concepts-with-a-Higgins-Matthey/a90226c41b79f8b06007609f-39f82757073641e2>
- [13] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," in *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, L. Rokach, O. Maimon, and E. Shmueli, Eds., Cham: Springer International Publishing, 2023, pp. 353–374. doi: 10.1007/978-3-031-24628-9\_16.
- [14] S. Zhao, J. Song, and S. Ermon, "Towards Deeper Understanding of Variational Autoencoding Models." arXiv, Feb. 28, 2017. doi: 10.48550/arXiv.1702.08658.
- [15] H. Zhao, P. Rai, L. Du, W. Buntine, D. Phung, and M. Zhou, "Variational Autoencoders for Sparse and Overdispersed Discrete Data," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, PMLR, Jun. 2020, pp. 1684–1694. Accessed: Dec. 11, 2023. [Online]. Available: <https://proceedings.mlr.press/v108/zhao20c.html>
- [16] C. Doersch, "Tutorial on Variational Autoencoders." arXiv, Jan. 03, 2021. doi: 10.48550/arXiv.1606.05908.
- [17] K. Gregor, G. Papamakarios, F. Besse, L. Buesing, and T. Weber, "Temporal Difference Variational Auto-Encoder." arXiv, Jan. 02, 2019. doi: 10.48550/arXiv.1806.03107.
- [18] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The Variational Fair Autoencoder." arXiv, Aug. 09, 2017. doi: 10.48550/arXiv.1511.00830.
- [19] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, Jun. 2015, pp. 2256–2265. Accessed: Dec. 13, 2023. [Online]. Available: <https://proceedings.mlr.press/v37/sohl-dickstein15.html>
- [20] J. Ho, A. Jain, and P. Abbeel, "Denosing Diffusion Probabilistic Models".
- [21] J. Betker et al., "Improving Image Generation with Better Captions".
- [22] R. Bommasani et al., "On the Opportunities and Risks of Foundation Models." arXiv, Jul. 12, 2022. doi: 10.48550/arXiv.2108.07258.
- [23] Gemini Team et al., "Gemini: A Family of Highly Capable Multimodal Models." arXiv, Dec. 18, 2023. doi: 10.48550/arXiv.2312.11805.

- [24] A. Vaswani et al., "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: May 22, 2023. [Online]. Available: [https://papers.nips.cc/paper\\_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html](https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html)
- [25] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018.
- [26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *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)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
- [27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multi-task Learners," 2019.
- [28] T. Brown et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 1877–1901. Accessed: Oct. 18, 2023. [Online]. Available: <https://papers.nips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- [29] H. W. Chung et al., "Scaling Instruction-Finetuned Language Models." arXiv, Dec. 06, 2022. doi: 10.48550/arXiv.2210.11416.
- [30] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning." arXiv, Sep. 02, 2021. doi: 10.48550/arXiv.2104.08691.
- [31] L. Ouyang et al., "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, Dec. 2022.
- [32] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model." arXiv, Dec. 13, 2023. doi: 10.48550/arXiv.2305.18290.
- [33] "Tools | 🦜 Langchain." Accessed: Feb. 12, 2024. [Online]. Available: <https://python.langchain.com/docs/modules/agents/tools/>
- [34] R. Yang et al., "GPT4Tools: Teaching Large Language Model to Use Tools via Self-instruction." arXiv, May 30, 2023. doi: 10.48550/arXiv.2305.18752.
- [35] T. Schick et al., "Toolformer: Language Models Can Teach Themselves to Use Tools." arXiv, Feb. 09, 2023. doi: 10.48550/arXiv.2302.04761.
- [36] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models." arXiv, Mar. 08, 2023. doi: 10.48550/arXiv.2303.04671.
- [37] P. Wang et al., "OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022, pp. 23318–23340. Accessed: Feb. 12, 2024. [Online]. Available: <https://proceedings.mlr.press/v162/wang22al.html>
- [38] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face." arXiv, May 25, 2023. Accessed: Sep. 11, 2023. [Online]. Available: <http://arxiv.org/abs/2303.17580>
- [39] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models." arXiv, Mar. 09, 2023. doi: 10.48550/arXiv.2210.03629.
- [40] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." arXiv, Jan. 10, 2023. doi: 10.48550/arXiv.2201.11903.
- [41] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior." arXiv, Aug. 05, 2023. Accessed: Sep. 12, 2023. [Online]. Available: <http://arxiv.org/abs/2304.03442>
- [42] Q. Wu et al., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework." arXiv, Aug. 16, 2023. doi: 10.48550/arXiv.2308.08155.
- [43] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, in NIPS'20. Red Hook, NY, USA: Curran Associates Inc., Dezember 2020, pp. 9459–9474.
- [44] "LangChain: A Complete Guide & Tutorial," Nanonets Intelligent Automation, and Business Process AI Blog. Accessed: Feb. 12, 2024. [Online]. Available: <https://nanonets.com/blog/langchain/>
- [45] "Prompt Engineering for Generative AI | Machine Learning," Google for Developers. Accessed: Feb. 12, 2024. [Online]. Available: <https://developers.google.com/machine-learning/resources/prompt-eng>
- [46] "Prompts | 🦜 Langchain." Accessed: Feb. 12, 2024. [Online]. Available: [https://python.langchain.com/docs/modules/model\\_io/prompts/](https://python.langchain.com/docs/modules/model_io/prompts/)

## Contact

---

Fraunhofer Institute for  
Applied Information Technology FIT  
Branch Business & Information Systems Engineering  
Wittelsbacherring 10  
95444 Bayreuth | Germany  
Phone +49 921 55-4710  
[info@fit.fraunhofer.de](mailto:info@fit.fraunhofer.de)  
[www.wi.fit.fraunhofer.de](http://www.wi.fit.fraunhofer.de)

Fraunhofer Institute for  
Energy Economics and Energy System Technology  
Joseph-Beuys-Straße 8  
34117 Kassel | Germany  
Phone +49 561 7294-0  
[empfang@iee.fraunhofer.de](mailto:empfang@iee.fraunhofer.de)  
[www.iee.fraunhofer.de](http://www.iee.fraunhofer.de)