

Operational Excellence

Market Insight: Ten Applications Of Large Language Models For Industry

By Joe Lamming
With Malavika Tohani

July 2023



Market Insight: Ten Applications Of Large Language Models For Industry

By Joe Lamming
With Malavika Tohani

July 2023

This report highlights the potential of large language models (LLMs), first brought to the mainstream by OpenAI's ChatGPT. It details their inner workings and outlines how they can be integrated as agents within current technologies, data structures and industrial procedures, to create substantial market impact. We explore the Transformer architecture, its significance, and its limitations. Drawing upon meticulous analysis of recent academic research, commercial product launches, and contributions from the open-source AI community, this report guides readers through comprehensive explanations and intuitive graphical depictions of LLMs functioning as chatbots, reasoning companions, information retrieval systems and agent-style task completion aides. Our findings illuminate a future in which data management, machine learning and LLMs intersect to fuel unparalleled efficiency and innovation in the heavy industries sector.

Table of contents

Advances in data management and machine learning techniques have become rocket fuel for AI development	3
Transformer-based LLMs dominate today's generative AI landscape	
Ten applications of LLMs for industrial operations	

Table of figures

Figure 1. The transformer architecture enables deep language understanding	4
Figure 2. Encoders distil the meaning of sentences	5
Figure 3. Transformer decoder LLMs generate text	6
Figure 4. LLMs can model and predict real-world cause and effect	7
Figure 5. Retrieval-augmented generation (RAG) can eliminate LLM hallucination	8
Figure 6. LLM agents can use tools to complete tasks	9

Organizations mentioned

ABB, Anthropic, Apple, AspenTech, AVEVA, C3 AI, Cognite, DataStax GitHub, Google, Hexagon, HighByte, Hitachi Vantara, International Electrotechnical Commission (IEC), Meta, Microsoft, NVIDIA, OpenAI, Rockwell Automation, Salesforce Research, Seoul National University, Siemens, SparkCognition, Stability AI, Technology Innovation Institute, ThirdAI, Timeseer.ai.



Advances in data management and machine learning techniques have become rocket fuel for AI development

Language, to some, is a game where the objective is communication. From the abacus of antiquity, to Charles Babbage's Difference Engine's knobs and dials in the early 19th century, to 20th century hole-punched paper – to today's text-based programming languages like JavaScript, C++, and Python – these advancements represent the continuing human drive towards accessible computation.

Generative artificial intelligence (AI), a popular catch-all term for highly sophisticated chatbots, image and video generators and speech synthesizers, has enjoyed a warm winter in the spotlight, and is currently basking in the summer sunshine (see [Verdantix Market Insight: Understanding The Rapidly Evolving Landscape Of Generative AI](#)). While disparate use cases for generative AI call for vastly different model architectures, all – so far, without exception – require gargantuan quantities of data only made possible in recent years. Large language models (LLMs) have dominated discourse in generative AI, driving popular applications such as Anthropic's Claude, Google's Bard, Microsoft's Bing Chat and OpenAI's ChatGPT. Under the hood, LLMs are often described as 'just predicting the next word', based on prior context. Search engines, smartphone keyboards and email software all provide auto-complete functionality, taking in the context of what has been written so far to provide a best guess for what comes next. LLMs, however, greatly increase the accuracy of these predictions, by learning to guess the next token (word or word fragment) in a sentence through trial and error with vast quantities of text scraped from the public internet, digitized books, academic papers and proprietary sources.

In 2018, OpenAI's first generative pre-trained transformer (GPT), a deep neural network machine learning (ML) language model, was trained on around 5GB of text. Subsequent models were trained on ever larger data sets, with GPT-2 trained on 40GB of text and GPT-3 on 570GB – equivalent to around 220 billion words. In 2022, Google's Pathways Language Model (PaLM) was trained on around 600 billion words. Arriving in 2023, Meta's LLaMA and Dubai's Technology Innovation Institute's Falcon LLM learned from approximately one trillion words – while Google's PaLM 2 saw three trillion. Not only have advances in data acquisition contributed to the rise of vast, high-quality data sets for developing powerful generative AI – but so have ML techniques, optimizations and data management (see [Verdantix Smart Innovators: Industrial Data Management Solutions](#)). LLMs, distilling the natural language of billions of people, symbolize the next milestone, fostering effortless communication between data scientists, executives and customers – and in industry, between facilities managers and frontline workers.

Transformer-based LLMs dominate today's generative AI landscape

In 2017, researchers at Google Brain, now part of Google DeepMind, released a novel deep learning architecture named the Transformer. Its ability to efficiently split inputs, such as text or images, into pieces, through a process called self-attention, paved the way for the development of powerful language translation, summarization and even computer vision systems. For natural language, this enables a model to understand the deeper meaning of a whole sentence. All of today's most powerful LLMs utilize variations of the transformer architecture.

Transformer-based LLMs enable:

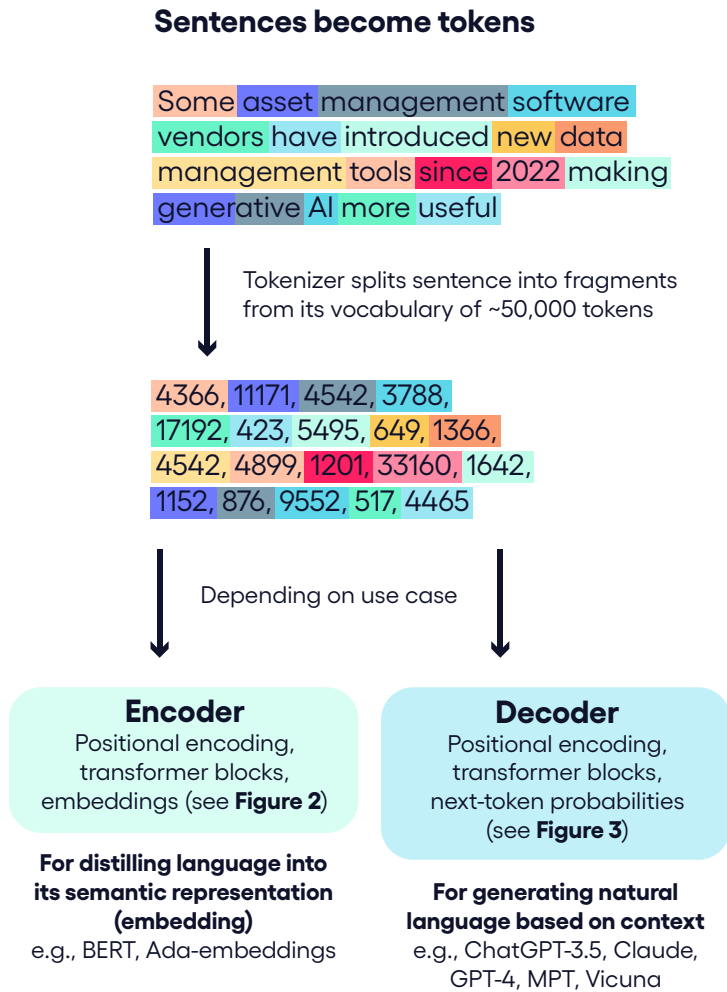
- **Deep understanding of natural language.**

In enterprise and industrial environments, the ability to quickly search large document stores has traditionally been provided by keyword matching and 'fuzzy search' algorithms. While such methods deliver results at high speed when looking for specific words, part numbers and personnel, there is a reason people use sentences to communicate complex information rather than single words. Words, in sequence, provide context to one another. Understanding intent is key to turning words into actions – and the transformer architecture's multiple attention heads help it understand not only individual words, but how they relate to other words, and therefore encode the deeper meaning of the entire sentence (see **Figure 1**). This ability to distil the deeper meaning of sentences – known as 'embeddings' – also enables LLMs to determine the semantic similarity of other sentences, through a process called 'vector comparison' (see **Figure 2**). Such embeddings can be based on images, allowing vastly more powerful industrial data search capabilities.



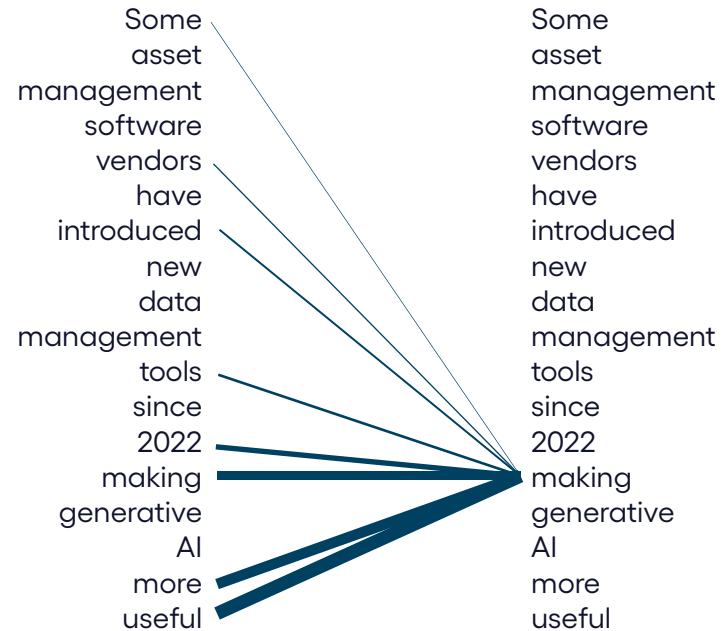
Figure 1

The transformer architecture enables deep language understanding



Self-attention extracts meaning

Every word in the sentence “**Some asset management software vendors have introduced new data management tools since 2022 making generative AI more useful**” is contextualized with every other word and itself. This helps the model extract meaning from the sentence.



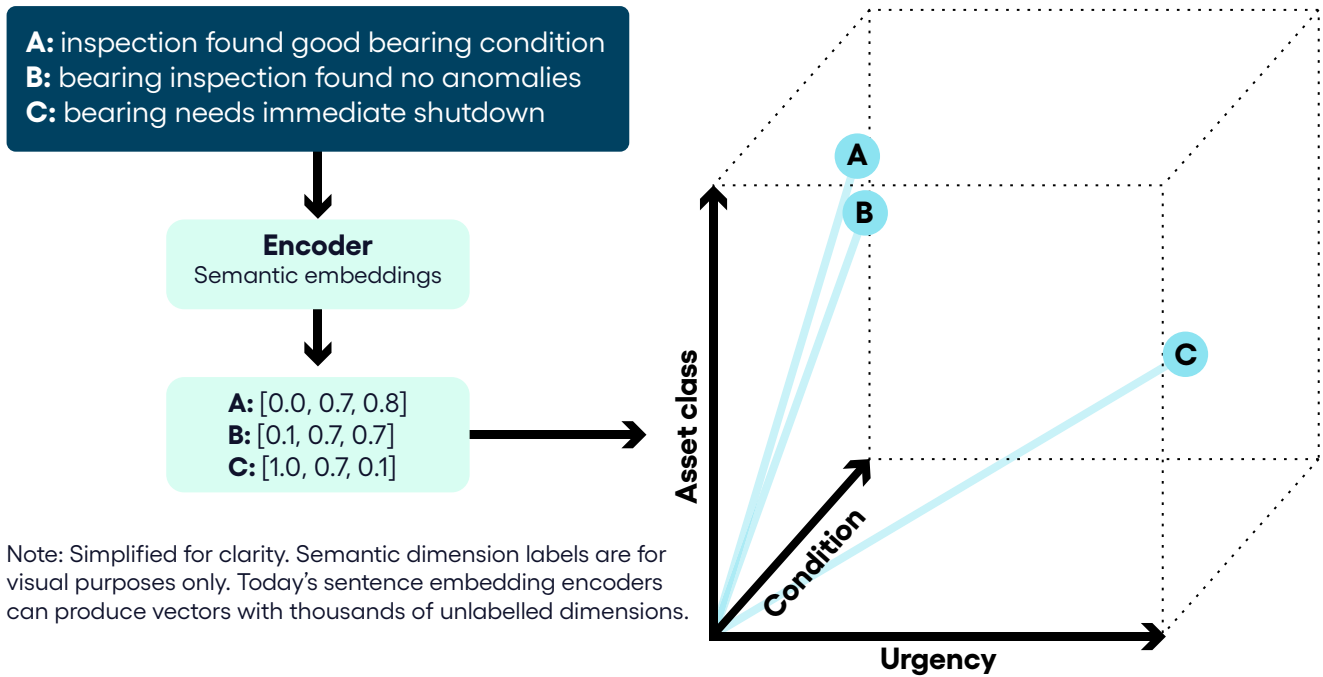
Note: Tokenization example performed using OpenAI GPT-3 tokenizer.

Source: Verdantix analysis, derived from 'Attention Is All You Need' (Vaswani et al., 2017).



Figure 2

Encoders distil the meaning of sentences



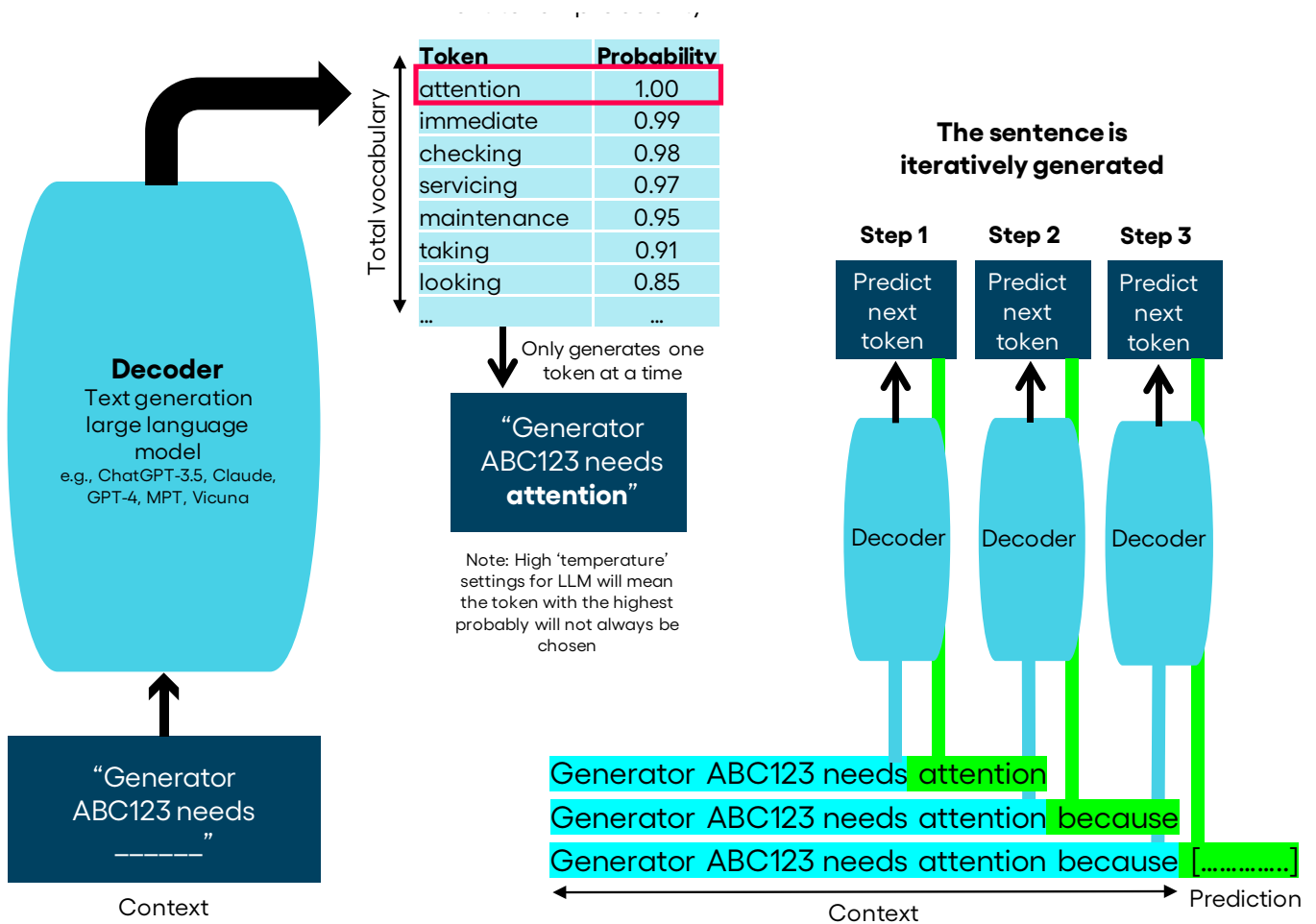
Each dimension **A**, **B** and **C** represents a semantic representation. By learning from billions of sentences, the transformer encoder can extract meaning and represent this as a multidimensional vector. Points **A** and **B** are close together because they have similar meaning.

Source: Verdantix analysis

- **Two-way, conversational interfaces.**

Popular web search solutions, such as Google Search, have performed transformer-based natural language processing (NLP) on queries to extract intent and serve relevant results since 2019. However, at present, conversational interfaces for users to refine web search queries, such as Google Bard or Microsoft Bing Chat, have yet to see mass adoption. For unusual queries, such as where the user is new to a topic and performing research for the first time, back-and-forth communication enabled by chatbot-style LLMs provide every user, industrial operative, business executive and data scientist with the ability to tackle complex problems at their own pace. Through a process called 'reinforcement learning from human feedback' (RLHF), the raw, uncontrolled next-token prediction abilities of foundational LLMs are fine-tuned to assume the role of a friendly assistant, making the model more stable, less toxic and easier to converse with. These types of LLMs, popularized by Google's LaMDA and OpenAI's ChatGPT in 2022, enable conversations through their ability to understand the back-and-forth nature of dialogue and to hold thousands of words in their context window.

Figure 3
Transformer decoder LLMs generate text



Source: Verdantix analysis

- **Understanding of causality beyond simple auto-complete.**

A next-word prediction system working within the context of only a few words has limited value. However, if such a system is allowed to iteratively generate an entire sentence, paragraph, or inspection report – and does so with full relevance to an initial context window containing a difficult problem – then we are in fact witnessing the LLM model reality itself (see **Figure 3**). While some sentence completions are easy, others require a deeper understanding of cause and effect. Today's state-of-the-art LLMs will, during training, attempt to predict the next word trillions of times – and, in doing so, become effective general reasoning engines with applications across consumer, enterprise and heavy industry. Such systems will excel when reasoning about topics they have seen extensively during training. However, they are also vulnerable to so-called 'hallucinations' – where the model will generate a reasonable-sounding but incorrect answer when presented with a novel situation or a query too complex for it to understand (see **Figure 4**).

Figure 4

LLMs can model and predict real-world cause and effect

Simple chat completion problem	Complex chat completion problem
<i>[Brief description of the industrial facility, the safety protocols and the role of the chatbot]</i>	<i>[Brief description of the industrial facility, the safety protocols and the role of the chatbot. Description of ability to shut off facility power or activate emergency alarm is also provided]</i>
Observation: A worker enters the facility. CCTV detects they are wearing work boots, blue overalls and sunglasses.	Observation: CCTV detects that Worker A is interacting with air compressor ABC123. The real-time monitoring system then detects that air compressor ABC123 power draw has fallen to zero and the facility pneumatic pressure is rapidly falling in Zone D. CCTV then detects Worker A disassembling the underside of the air compressor, with some components removed. The worker then walks away from the compressor. A few minutes later, CCTV detects that Worker B is interacting with the pneumatic table lift in Zone D and appears to be frustrated. Worker B approaches air compressor ABC123 and interacts with it. The real-time monitoring system then detects that air compressor ABC123 power draw has increased to 212% of maximum rated power and the facility pneumatic pressure for Zone D is increasing at the rate of 0.5 bar per second and is rated for maximum of 10 bar. The facility manager will be unavailable for the next 10 minutes.
Action [correct chatbot response]: Based on the observation, it appears that the worker is not wearing the complete required personal protective equipment (PPE) as per the facility's safety protocols. The worker is wearing work boots and blue overalls, which are appropriate for the work environment, but they are missing a hard hat, reflective vest, and safety glasses. The sunglasses may not provide adequate eye protection. To address this situation, I would recommend the following actions: <ul style="list-style-type: none">Identify the workerCommunicate with the workerProvide PPETraining and reinforcementCompliance monitoringIncident reporting	Action [incorrect chatbot response]: Based on the observations, it appears that there is a potential risk incident involving air compressor ABC123 and the pneumatic pressure in Zone D. Here are the recommended actions to address the situation: <ul style="list-style-type: none">Alert Worker A and Worker B using the facility's intercom system.Evacuate Zone D Correct action: Shut off facility power.

Source: Verdantix experiments with OpenAI ChatGPT-3.5

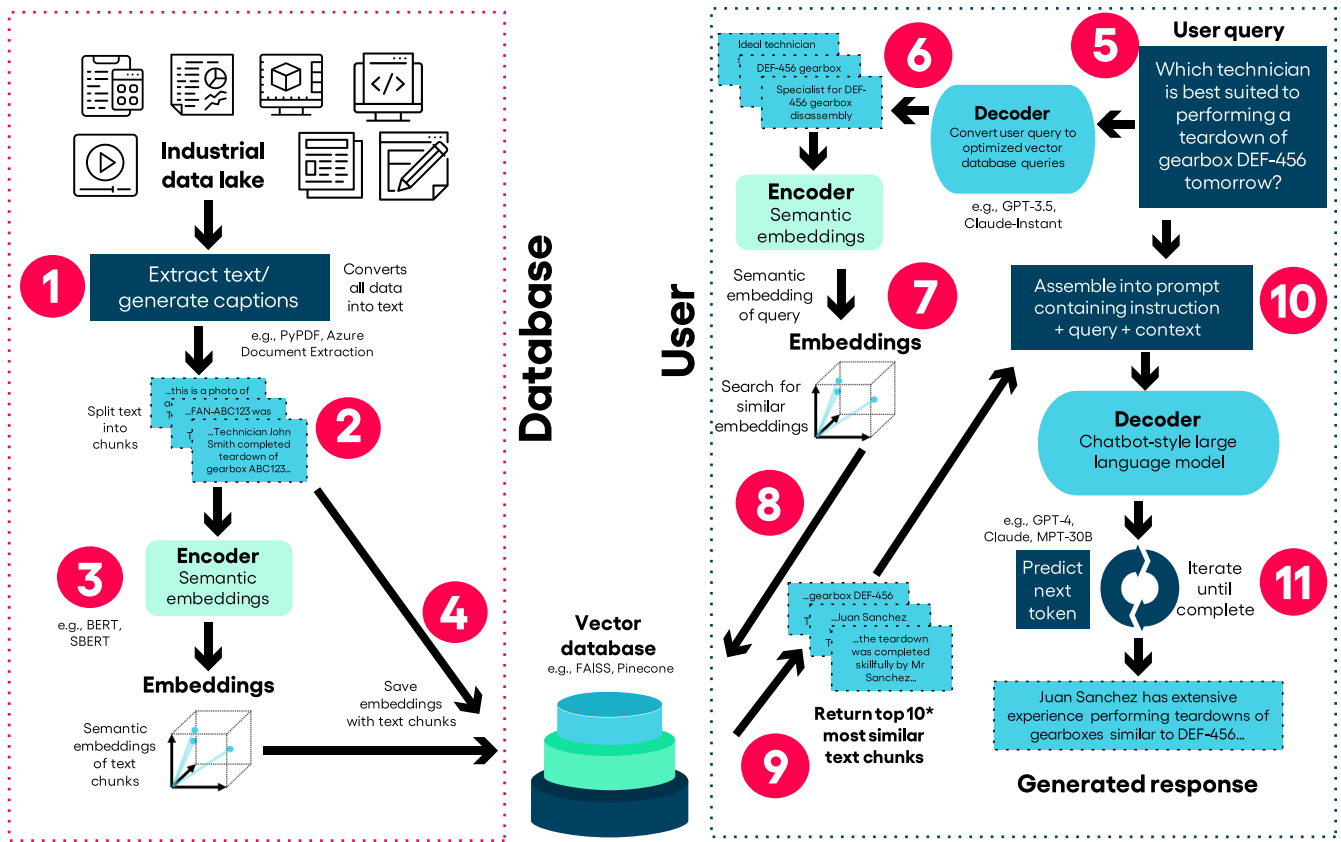
- **Generation of creative language, but not hard facts.**

Generative language models are neural networks, not databases. LLMs alone are not designed to retrieve facts and regurgitate them when asked to – instead, they simply produce a probability map across approximately 50,000 tokens (words) in their vocabulary, based on their original training data and the context already provided in the prompt. Maintaining absolute truth in vast training data sets is infeasible with even today's best data management solutions. The development of the most powerful LLMs is highly resource-intensive, making continuous re-training equally impractical. Meta's largest LLaMA-65B model took 21 days of training with 2,048 NVIDIA A100 graphics processing units (GPUs) – around \$25 million worth at the time – an impossible cost and timeline for most industrial firms. Larger models such as Google's latest PaLM 2 are rumoured to cost hundreds of millions. Attempts to avoid such costs by fine-tuning models only on new data, or with smaller data sets, such as a firm's internal documents, can result in degraded performance on general tasks – hurting the LLM's abstract reasoning abilities. However, when deployed to parse, summarize and rephrase text provided in context, LLMs are valuable digital assistants.



Figure 5

Retrieval-augmented generation (RAG) can eliminate LLM hallucination



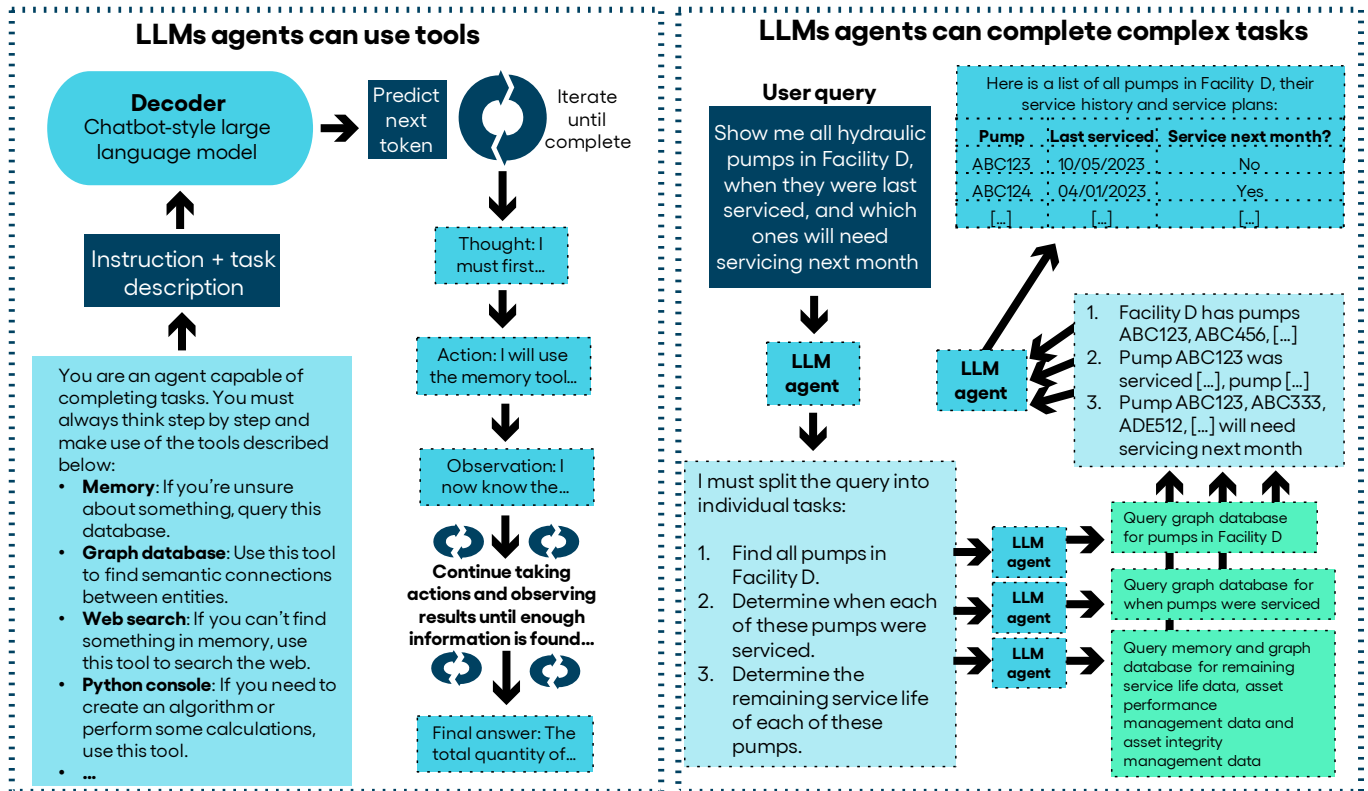
Note: Optimal number of chunks returned is determined by LLM context limit, RAG prompt design and other factors.
 Source: Verdantix analysis

- Powering of sophisticated information retrieval and synthesis.**

To make use of large data sets, LLMs need external pipework. The architecture of today’s best transformer neural networks means that the maximum amount of context that can be analysed by the model has a significant effect on its computational cost. If the context length is doubled, the model’s computational requirements are quadrupled. Meta’s LLaMA has a context limit of 2,048 tokens (~1,500 words), while OpenAI’s latest June 2023 GPT-3.5-Turbo has ~16,000; the latter’s limited-release GPT-4-32K can handle ~32,000 tokens, while Anthropic’s Claude can take in ~100,000 (~75,000 words). With today’s most popular LLMs, forcing them to ingest more context than they were originally trained on will produce erroneous generations. However, advances up to June 2023 have enabled context to be effectively increased up to 16 times by lightly fine-tuning existing models, such as through Meta AI’s Positional Interpolation of LLaMA. Similarly, breakthroughs in the more mature retrieval-augmented generation (RAG) technique are already enabling existing LLMs to utilize ‘chunks’ of semantically relevant text retrieved from a database of embeddings, to provide context to generated text (see **Figure 5**). Further improvements to RAG systems are being brought by innovations in highly efficient, continually training retrieval, such as the NeuralDB by start-up ThirdAI. In May 2023, ThirdAI partnered with DataStax to deliver its BOLT AutoML solution to enterprise customers at scale.



Figure 6
LLM agents can use tools to complete tasks



Source: Verdantix analysis, derived from papers on LangChain, AutoGPT and Recursion-of-Thought (RoT) by Soochan Lee, Seoul National University

- Completion of tasks through divide-and-conquer and recursion-of-thought.**

In the same way that humans can accidentally misremember a fact or make a basic error when performing complex mental calculations without first writing out the problem step by step, LLMs may often incorrectly predict an answer or make a false assumption on the first try. Due to their auto-regressive nature, these initial falsehoods generated by the LLM will persist and affect future text. While workarounds involve prompting the model to reduce complexity by thinking step by step, this is still error-prone. Solutions include LLMs using external tools such as calculators, software development consoles, and database and search engine application programming interfaces (APIs) to gather niche and/or up-to-date information to complete specific tasks (see **Figure 6**). In June 2023 OpenAI released 'function calling' improvements to its GPT-3.5 and GPT-4 APIs to improve the reliability of these LLMs when interacting with external tools through JSON format chat completions. In June 2023 Cognite launched a generative AI product offering a low-code interface to the full functionality of its DataOps platform – Cognite Data Fusion – to reduce reliance on the distilled general knowledge of LLMs and instead make use of carefully curated industrial data.



Ten applications of LLMs for industrial operations

The characteristics and capabilities of transformer-based LLMs are increasingly well understood – such as unrivalled understanding and masterful generation of human language, and a friendly tendency towards conversational instruction-following. Similarly, stronger LLMs such as GPT-4 and Claude exhibit remarkable understanding of real-world cause and effect, with the former rumoured to be composed of eight GPT-3.5-sized LLMs in a mixture-of-experts (MoE) configuration. Limitations, however, present as a tendency to hallucinate facts, malfunctions when provided with too much context, and failures at arithmetic – although these are already seeing solutions in the form of careful prompting, RAG and wraparound software to enable agent-like behaviour. In the words of OpenAI CEO Sam Altman, these all represent a “Cambrian explosion” of possibilities. Simultaneously, these possibilities will increase pressure on regulators to take action, such as with the EU’s 2021 AI Act (see [Verdantix The EU Sounds The Clarion Call On Artificial Intelligence Regulation](#)). Meanwhile, operations, maintenance and process safety executives are under intense pressure to optimize production, increase yield, reduce emissions and meet increasingly stringent safety standards. In the midst of tension between rapid technological evolution, regulation and societal worries, Verdantix sees ten high-value use cases of generative AI for industry. These are:

1. Extracting relevant critical information from vast data sets for concise insights.

As digitization is rolled out across industrial enterprises, the resulting data warehouses and data lakes will store everything from decades of high-frequency sensor measurements across thousands of Internet of Things (IoT) devices, to millions of inspection reports, work orders, scanned notes and production logs. Powerful image captioning tools, such as BLIP-2 by Salesforce Research, enable the enrichment of visual data with text-based metadata, while table and document parsing tools by firms such as C3 AI and Cognite offer LLMs visibility into multimodal data. By employing retrieval systems to serve text chunks to LLMs, operators are provided with conversational, grounded-in-truth representations of relevant data (see **Figure 5**). Cognite’s Industrial Knowledge Graph provides LLMs with semantic relationships between assets, processes, technologies and people, to reduce hallucinations. LLM-based information retrieval systems give operators concise, relevant insights for a big-picture view – helping them discover inefficiencies and safety risks.

2. Eliminating repetitive administrative tasks through automation.

Technologies such as digital twins, AI analytics and asset management software help automate multiple processes at industrial facilities, with 87% of the 301 respondents in the 2022 Verdantix global corporate operational excellence survey mentioning the availability of new technologies as the most significant factor driving digital transformation of plant operations (see [Verdantix Global Corporate Survey 2022: Operational Excellence Budgets, Priorities & Tech Preferences](#)). LLMs will enhance these capabilities even further by performing mundane, repetitive administrative tasks such as drafting emails, scanning reports to triage risks and retrieving information from systems where conventional software integration has not been implemented. In April 2023 Siemens announced a collaboration with Microsoft to launch its new Teamcenter app within Microsoft Teams, helping shop floor workers parse and translate natural speech, generate summarized reports and route information to appropriate design, engineering or manufacturing personnel.

3. Enabling more robust industrial data ingest, transformation and contextualization.

Industrial data can be vast, inscrutable and expensive to manage without suitable tools (see [Verdantix Strategic Focus: Why Industrial Firms Need DataOps Platforms For Asset Management Digitization](#)). Firms such as AspenTech, AVEVA, HighByte and Hitachi Vantara offer industrial DataOps platforms to meet diverse data management needs, while others, such as Timeseer.ai, provide specific tools to detect and provide alerts for more than 100 data quality issues. LLMs excel at parsing unstructured data, using reasoning to add context, and troubleshooting software issues. Deployed as agents, generative AI will greatly increase the ease of use of data management and orchestration (see **Figure 6**). Included in Cognite’s Industrial Canvas platform is multimodal contextualization within a single pane of glass view, powered by LLM-based agents and generative AI.



4. Offering ops & maintenance workers a quick second opinion by acting as a reasoning engine.

The ability of RLHF-tuned LLMs to follow natural language instructions allows them to explore their digital environment through chain-of-thought or recursion-of-thought reasoning in a way that is understandable to humans. They can query industrial data lakes, read and summarize documents, or review real-time data through connections to enterprise asset management (EAM), EHS or asset performance management (APM) software. Deployed as agents – to perform a task based on user instructions – LLMs can undertake much of the mundane knowledge-gathering and basic analysis, streamlining frontline worker tasks such as fetching a list of specific assets (for example, pumps) in a facility, noting their service history, and predicting which ones will need servicing next month (see **Figure 6**). While even today's most powerful LLMs, such as GPT-4 and Claude, will sometimes make mistakes, their general knowledge of the world, when utilized with the appropriate software scaffolding to direct their attention, offers operators, managers and engineers a quick, judgement-free sanity check or second opinion on critical decisions (see **Figure 4**).

5. Automatically categorizing and prioritizing asset maintenance tasks.

Expert at analysing unstructured data – either directly, from text scraped from documents, or from captions generated by AI vision models – LLMs have limitless patience to continuously monitor real-time information uploaded to industrial data lakes. Such functionality can be leveraged to extract sentiment from data, compare it with operational priorities and serve summaries to facilities and corporate decision-makers accordingly. Similarly, LLMs can use risk and criticality metrics to screen thousands of inspection reports, image captions and available transcripts from calls to detect imminent incidents and provide timely alerts through agent-style process automation to site managers.

6. Facilitating fully hands-free operation with voice dictation for inspections and maintenance.

For more than a decade, dictation has been deployed on smartphones in the form of Apple's Siri, Google's Assistant and numerous others. However, such systems have been limited in their ability to recognize domain-specific words or consistently extract complex instructions. In 2022 OpenAI released the open-source Whisper model – a versatile, general-purpose speech-to-text system trained on 680,000 hours of transcripts. Such models can be combined with LLMs and vision systems to feed a virtual assistant and provide audio and visual information to operators in the field, hands-free. While Whisper and similar models are currently computationally expensive, enterprise-focused value from accurate transcription is driving innovations and enabling the rapid development of compact models trained to recognize industry-specific terminology. Such systems will offer frontline workers a software-based reasoning engine and virtual assistant to help with complex tasks, especially in remote locations.

7. Democratizing asset programmable logic controller (PLC) programming.

The rigorous logic required by computer programming languages, alongside the ubiquity of thorough discourse around software development on the internet, means that LLMs have learned to closely associate code with natural language. In the industrial space, machine vendors such as ABB, Rockwell Automation and Siemens offer extensive public documentation for programming their products. Microsoft-owned GitHub Copilot, launched in 2021 and widely available from 2022, offers sophisticated auto-complete features to software developers, including the ability to generate a function based on a natural language description. Similarly, in May 2023 ABB Research published a paper detailing how OpenAI's ChatGPT/GPT-4 uses natural language description of PLC/DCS functionality to generate syntactically correct IEC 61131-3 Structured Text code and demonstrate useful reasoning skills to boost control engineer productivity, alongside control narratives.

8. Delivering a low-code, natural, conversational interface to the whole workforce.

LLMs offer a universal translation layer between global human languages and the code-heavy or domain language used across industrial databases and software solutions. Today, many industrial software solutions rely upon carefully designed graphical user interfaces (GUIs), application-specific layouts and extensive user training programmes to help customers get the most value out of them. However, problem-solving in



the field can require the use of disparate tools and software solutions – where GUIs are too restrictive and interoperability between competing vendors is limited. The ability of LLMs to utilize code-level interfaces through common programming languages such as Python allows users to leverage the powerful granular functionality of a platform. In June 2023 Hexagon launched its HxGN EAM Python Framework, while C3 AI offers code-level functionality through its Type System for a variety of programming languages. Also in June, Cognite launched its Copilot product to utilize the natural communication abilities of LLMs to act as a general-purpose low-code interface to its solution’s most advanced features – thereby providing far more frontline workers, data scientists, facilities managers and executives with the ability to interact with critical information through their preferred medium.

9. Developing more advanced AI-based vision systems for production quality optimization.

Computer vision and LLMs were distinctly different technologies up until 2020, when the vision transformer (ViT) model deployed the architecture designed for language to analyse a sequence of image patches, to better understand visual data. In 2021 OpenAI’s CLIP model utilized the ViT to recognize complex visual features, while in June 2023 Salesforce Research’s BLIP-2 deployed a CLIP-based ViT combined with an LLM to allow conversational interaction with images. By fine-tuning vision models and LLMs to provide domain-specific insights, quality management on production lines will see improved accuracy and domain expert skills will be better utilized. Other vision-based models can help fill in missing data. In May 2023 SparkCognition announced a collaboration with Shell to deploy image-based generative AI to shorten the time required to conduct seismic surveys from nine months to just nine days.

10. Providing richly visual 3D virtual environments for training.

Latent diffusion models, such as OpenAI’s DALLE-2 and Stability AI’s Stable Diffusion, generate convincingly realistic surroundings based on sparse natural language prompts. Open-source projects utilizing derivations of Stable Diffusion, such as ControlNet, offer fine controls over the generation of images – including the ability to enrich virtual environments with domain-specific scenery. Another project, NVIDIA’s NeuralField-LDM, uses hierarchical latent diffusion models to generate realistic, complex 3D scenes. Such technologies will give frontline workers unprecedented access to immersive training environments, helping with knowledge transfer and reducing operational risk.



Independent insight and analysis

Our research is a trusted source for some of the largest and most innovative businesses in the world. With over a decade of reports, data and analysis, our subscribers have access to depths of insight that cannot be found elsewhere.

Whether you are implementing a leading-edge technology strategy, or developing the products and value propositions of the future, our analysis will help you futureproof your thinking.

Our expertise

Environment, Health & Safety

ESG & Sustainability

Net Zero & Climate Risk

Operational Excellence

Smart Buildings

Contact

Verdantix Ltd, 18 Hatfields, London
SE18DJ, United Kingdom

contact@verdantix.com
[@Verdantix](https://www.verdantix.com)

Opportunities at Verdantix

Since 2008, Verdantix has been delivering high-quality research and advice to its clients. If you're interested in joining a world-class team with an unwavering focus on success, apply to join us today. We are delighted to be hiring across all teams and have a variety of opportunities in both London and Boston

