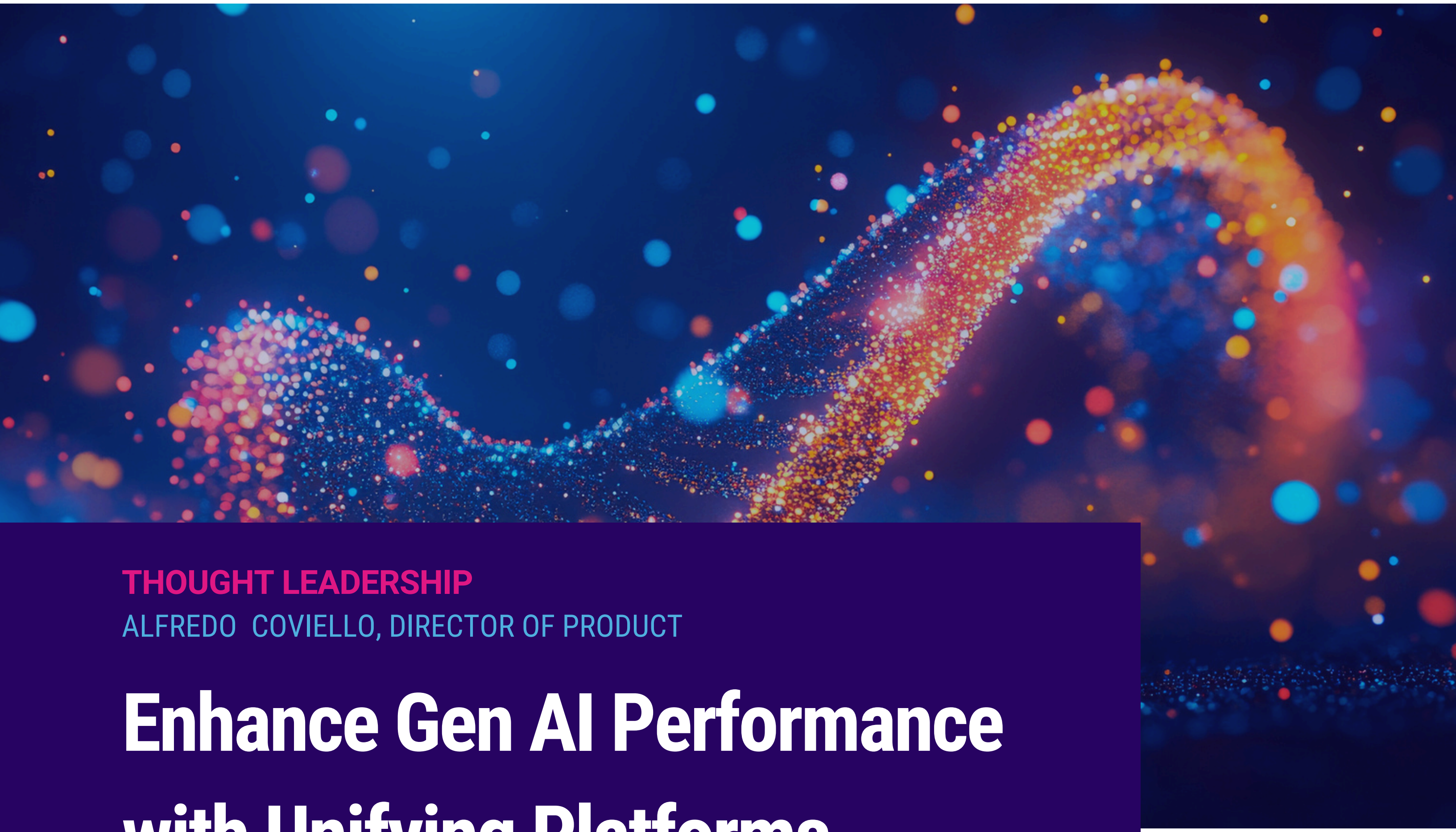
L7 INFORMATICS



THOUGHT LEADERSHIP

ALFREDO COVIELLO, DIRECTOR OF PRODUCT

Enhance Gen AI Performance with Unifying Platforms

In the rapidly evolving landscape of artificial intelligence (AI), particularly in generative AI (Gen AI), the quality and structure of data play a critical role. Gen AI models, known for their capabilities to generate human-like text, images, and other media, rely heavily on the data they are trained on. This article explores the importance of data quality, contextualized data, and data models in enhancing Gen AI performance and how the L7|ESP unifying platform provides a comprehensive ecosystem to boost these applications, with emphasis on its capabilities on data model management, knowledge graph, normalized schema, and ontology integrations. It also highlights how its regulatory-compliant orchestration engine lays out the foundation for successful Agentic AI implementations.

GENERATIVE AI USE, LIMITATIONS, AND SOLUTIONS

By now, many people use Gen AI in their daily lives for very different purposes. Gen AI models' transformative power comes from the vast amount of data they are trained on, the generic nature of applications they can be used for, and the speed with which they can produce somewhat expert, human-like responses.

Generative AI foundation models, most notably Large Language Models (LLMs) like GPTs (Generative Pre-trained Transformers), are a type of models that try to predict, or "inference", an output based on a user input. The input and output in Gen AI models are generally text (unimodal), but there are models that can accept or produce multiple forms, like images, audio, and video (multimodal).

Context Window Limitation

At its core, an LLM takes a piece of text and returns another, predicted piece of text based on the input. The number of words, or tokens, in the text is important because each model has a maximum number of tokens they can process, called the "context window", and because they are used to calculate usage cost. You can think of this window as the "attention span" window of a model, that is, the amount of data they can use in the input to generate an output.

Given this context window limitation, using LLMs with specific goals often revolves around optimizing interactions. A popular method is "prompting," which involves prefacing your interaction with a set of instructions about expected outcomes, roles to be played, and constraints. This is all about directing the model's focus and attention to the relevant aspects to provide the best possible result. In other words, the models perform better when adequate and relevant context is provided for the task at hand.

Training Dataset Limitation and RAG

LLMs are trained on diverse, curated, and generally publicly available datasets, which makes them ideal for applications across multiple fields. However, tasks requiring data not included in the training set—due to unavailability, selection criteria, or lack of public access—can render the model less effective (note that the training dataset cutoff date is usually clearly stated for LLM models; for example, GPT-4o is October 2023). A very common application of LLMs is using a model where private and confidential data needs to be used, like unpublished research or a company's proprietary documentation.

Retrieval-Augmented Generation (RAG), is a technique used for enhancing LLMs by incorporating private or more recent datasets. In a nutshell, the user query triggers fetching data and/or documents deemed relevant to the query and sending them along with the original query to the LLM. This allows the application to leverage all of the power of the model and its expansive training dataset while integrating crucial domain-specific and private insights.

Given the context window limitation, RAG must be handled carefully. If too much supplementary data is provided, the model may not use it properly, leading to incomplete or incoherent results. Conversely, insufficient data may yield suboptimal results.

The additional data provided by RAG to the model can be in different formats and be unstructured, semistructured, or structured. Structured data could come from a database, e.g., relational or graph. The ability to provide a knowledge graph (GraphRag) along with or instead of the unstructured text data sources is showing great promise in improving model performance. Note that an LLM could be used to actually generate a knowledge graph for a specific piece of text.

Model Chaining and Agentic AI

A common approach in AI and Machine Learning (ML) is chaining models, where the output of one model serves as the input of another. For example, an LLM generates the knowledge graph of a piece of text that is then used in a later query. This method allows the creation of complex and high-value outputs by chaining specialized applications.

Consider a diagnostic workflow as an example: An LLM gathers patient symptoms and clinical data, which is then input into a diagnostic model to suggest possible conditions. Based on current medical guidelines, an LLM may recommend additional tests, and a decision-support model might offer treatment options.

Each step of the process could use a specific model best suited for the task, with specific LLM prompts and associated RAG datasets and formats. In the example above, an image-to-text model could be added to provide suggestions for diagnostics based on patient imaging and the latest medical guidelines provided by RAG.

It's a common pattern in AI and ML to chain specialized and fine-tuned models in this way, in a known and tested workflow.

A top strategic trend identified by Gartner for 2025 is Agentic AI. The main idea is to use AI to actually determine the best workflow and actions for a specific task and goal at hand. The AI determines the workflow as the first step and is provided with a number of possible actions or tools it can use to execute the steps in that workflow, autonomously or semi-autonomously.

A common use case in LLM Agentic AI is literature review and update. Given a specific field of research, e.g. Alzheimer, and a concrete goal like "provide updates on novel findings", an agent could search and extract relevant articles by using a search tool and scientific database access, identify key topics and gaps in existing literature, update its knowledge base with new insights from latest publications, and generate periodic reports on emerging trends or potential breakthroughs. The key here is that the agent is provided with a concrete goal and is given specific tools that it uses with "autonomy" to achieve that goal.

For several reasons, there is an obvious need to supervise the activity of such agents, where an expert human-in-the-loop at high-value points of the chain can provide required approvals or even guidance before moving on in the list of tasks the agent executes. But this framework is expected to make impactful transformations in all fields, including Life Sciences, with increased productivity at the very top of the list.

As with any large-scale automation, ensuring that Agentic AI operates ethically, complies with regulations, and maintains security and safety standards is essential. This supervision guarantees the integrity of the AI's actions within the given framework.

THE DIFFERENT APPLICATIONS OF GEN AI IN AN ENTERPRISE LIFE SCIENCES SETTING

The potential for Gen AI applications in the life sciences is vast and will be transformative. We hear about them every day. Here are some key areas that are hugely benefiting from Gen AI applications:

Drug Discovery/BioTech

Generating molecular structures and predicting their properties to accelerate drug discovery processes.

- A cutting-edge AI research lab program uses AI to predict protein structures from amino acid sequences with remarkable accuracy.
- Another example is a biotech company that uses AI for drug discovery. It employs its generative adversarial networks (GANs) to design novel molecules with desired properties.
- One more example is a tech-driven pharma company, that uses automated wet labs and dry labs to collect, model, and analyze fit-for-purpose data and perform in silico experiments. Their machine-learning models identify the most promising targets, which are continuously improving. They have recently combined with another innovative drug discovery firm.

Clinical Drug Development

Pharmaceutical industry acceleration of clinical drug development.

- A large international pharma company is using ML and AI to improve clinical drug development by predicting

drug efficacy and side effects. It's also using AI to manage the vast amounts of generated data to produce faster and more accurate drug labels and regulatory documentation.
- Another global pharma company and a biotech firm are collaborating to develop protein therapeutics for multiple disease areas. The goal is to create protein-based drugs more quickly and cost-effectively than traditional methods.

CDMOs (large Contract Development and Manufacturing Organizations)

Optimization of biopharmaceutical manufacturing processes.

- A global contract development company uses AI to enhance process control, predict outcomes, and maintain quality across global production facilities
- Another large manufacturing services provider employs AI and machine learning to streamline the drug development and manufacturing process. Their AI-driven data analytics help accelerate the development timeline, improving quality and reducing costs.

Post-Market Surveillance

Analyzing masses of raw data and extracting insight into ADRs (Adverse Drug Responses) and other anomalies. Gen AI can also help with preparing documentation, communicating with patients and physicians, and updating marketing and packaging materials.

- A healthcare analytics company uses AI to analyze electronic health records, claims data and other real-world data sources to assess drug safety and effectiveness in the post-market phase.

Diagnostics

Improving diagnostic decision-making and patient treatment outcomes using patient data, medical literature, and medical records data (i.e., past medical history).

- A diagnostics technology company leverages ML to improve pathology diagnoses, aiming for more accurate and timely detection of diseases like cancer.

To develop these applications, one must start with the data.

THE IMPORTANCE OF DATA QUALITY FOR ML AND AI MODEL TRAINING

The foundation of any ML or AI model is the data it is trained on. Only high-quality data ensures that models learn the right patterns and make accurate predictions. Siloed data, that is disconnected data, and poor data quality, characterized by missing values, inconsistencies, and errors, can lead to biased or incorrect outputs.

In the context of Gen AI, where the objective is often to generate coherent and contextually accurate content, the importance of data quality cannot be overstated. The same applies to applications developed on top of those models, like RAG and model chaining.

> Clean, well-annotated, and diverse datasets are essential for training robust Gen AI models that perform well across various tasks and domains, including in scientific processes and data management.

THE IMPORTANCE OF DATA MODELS FOR ENHANCING AND ENSURING DATA QUALITY

Data models are structured frameworks for organizing and managing data. They help ensure data quality by providing consistency, reducing redundancy, and enabling efficient data retrieval. Data models define how data is stored, accessed, and manipulated within a database or data management system. Data Models are data about the data, enabling data to be described, understood, consumed, and transformed efficiently.

> Well-defined data models can lead to better-prepared training and inference datasets, streamlined data preprocessing, and more effective data integration processes, resulting in improved Gen AI models performance.

Gen AI performance is very much influenced by the type of data model used, including:

- **Enhanced Semantic Understanding**: By organizing data into meaningful structures, data models help Gen AI systems understand the relationships and context within the data, leading to more accurate and contextually relevant outputs.
- **Efficient Data Access and Retrieval**: Well-designed data models ensure efficient access to large datasets, reducing latency and computational overhead during model training and inference applications, like RAG.
- **Consistency and Integrity**: Data models maintain consistency and integrity, which are crucial for training models and their applications that generate reliable and factually correct information.

Additionally, leveraging structured data to improve LLMs performance is an active area of research and has shown promise in improving aspects of model performance, such as accuracy and comprehensiveness in data-rich domains.

THE IMPORTANCE OF CONTEXTUALIZED DATA IN LLM APPLICATIONS

Contextualized data, which refers to data enriched with context-specific information, is crucial for the efficacy of LLM applications. Contextualized data ensures that LLMs understand not just the content but the surrounding circumstances and nuances, leading to more accurate and relevant outputs. For example, a simple mention of a drug without context in a life sciences setting could lead to incomplete or incorrect interpretations. However, contextualized data would provide additional details such as dosage recommendations, potential side effects, interactions with other medications, and patient demographics. This enriched context allows LLMs to generate outputs that are not only factually correct but also tailored to the specific needs of the application, thereby significantly enhancing the reliability and usefulness of the generated content.

In the context of building the LLM models themselves, numerous academic papers and studies have specifically targeted the comparison between contextual and non-contextual models. Context and long-range dependencies are among the keys in this seminal paper, which is the foundation of most current language model architecture. Studies have shown that models like BERT and GPT achieve higher accuracy and F1 scores across multiple NLP tasks compared to models using Word2Vec or GloVe, which use static embeddings. Likewise, benchmarks like Glue and SuperGlue show contextualized models outperforming non-contextualized ones.

When using an LLM for any task, providing the right context is essential, given the context window size constraint, especially for complex tasks. Prompt engineering and other techniques have emerged to improve LLM's performance, which focuses on bringing attention to what is relevant to the task at hand, and it is all about its context.

Figure 1 illustrates the multifaceted nature of managing sample data. Each segment around the central "Sample" underscores the varied contexts in which data exists. These contexts are pivotal in ensuring that AI models generate relevant and accurate information. Let us delve into how contextualized data is essential for each aspect depicted in the image:

Inventory: Knowing the availability, storage conditions, and quantity of samples is critical for managing resources efficiently. Contextualized data ensures that AI models can predict when new samples will be needed, or when existing samples are approaching expiration.

Registration & Identification: Accurate identification and registration are foundational for traceability. Contextual data ensures that each sample is correctly identified, linked to its origin, and tracked through its lifecycle, thus maintaining the integrity of downstream analyses.

Tracking/Storage: Accurate identification and registration are foundational for traceability. Contextual data ensures that each sample is correctly identified, linked to its origin, and tracked through its lifecycle, thus maintaining the integrity of downstream analyses.

Labeling: Proper labeling with contextual data such as sample type, collection date, and handling conditions ensures that samples are used appropriately in experiments, reducing the risk of cross-contamination or misuse.

Analytical Data: Analytical results are meaningful only when contextualized with sample metadata. This includes experimental conditions, instruments used, and calibration details, which are critical for replicating and validating results.

Experimental Characterization (EXP1 & EXP2): In experiments, knowing the protocol, reagent details, and environmental conditions contextualizes the results, allowing AI models to generate accurate interpretations and predictions based on experimental data.

Planning & Scheduling: Effective planning and scheduling of experiments rely on knowing the availability of samples, equipment, and personnel. Contextualized data ensures that schedules are realistic and optimized, reducing downtime and resource conflicts.

Additionally, free-form, unstructured data, such as text annotations and graphs, can be associated with and combined with these datasets to capture even more relevant context, such as a patient Notebook or a Design of Experiment Notebook.

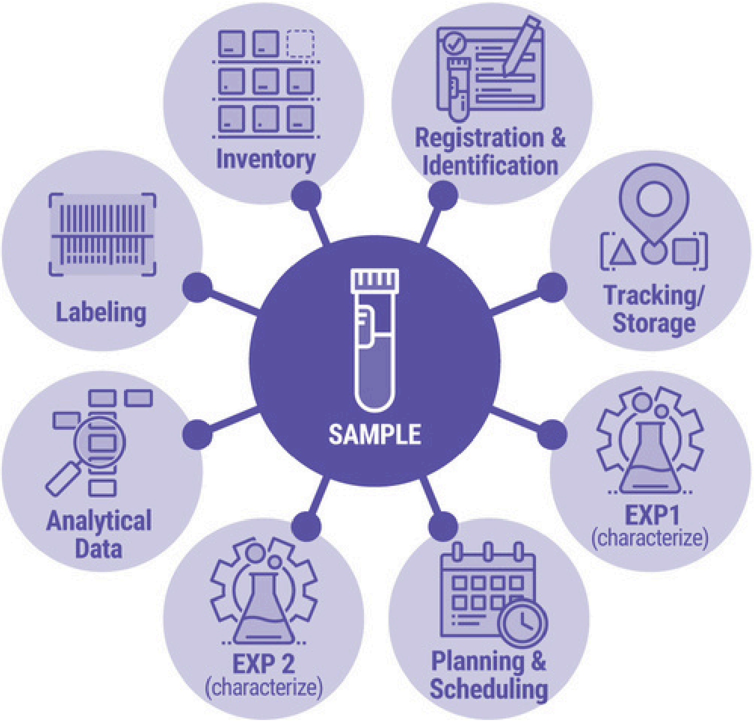


Figure 1: Type of contextualized data collected during the orchestration of laboratory and business processes as demonstrated for laboratory sample.

HOW DOES L7|ESP BOOST GEN AI PERFORMANCE?

L7|ESP is designed to significantly enhance the performance of Gen AI models through its robust data management and integration capabilities.

Data Model Management

L7|ESP gives its customers full control of the models to suit the use case at hand. The structured data created from those models is ready for use in LLM applications. This structured data can include star schema data products built using L7|INTELLIGENCE, which can be tailored to specific segments and verticals.

L7 also provides Reference Models that can be used out of the box but can also be fine-tuned and extended for specific needs. The models encompass and integrate the whole value chain including Discovery and Research, Development, Diagnostics, FDA submission, Manufacturing, and Post-market monitoring.

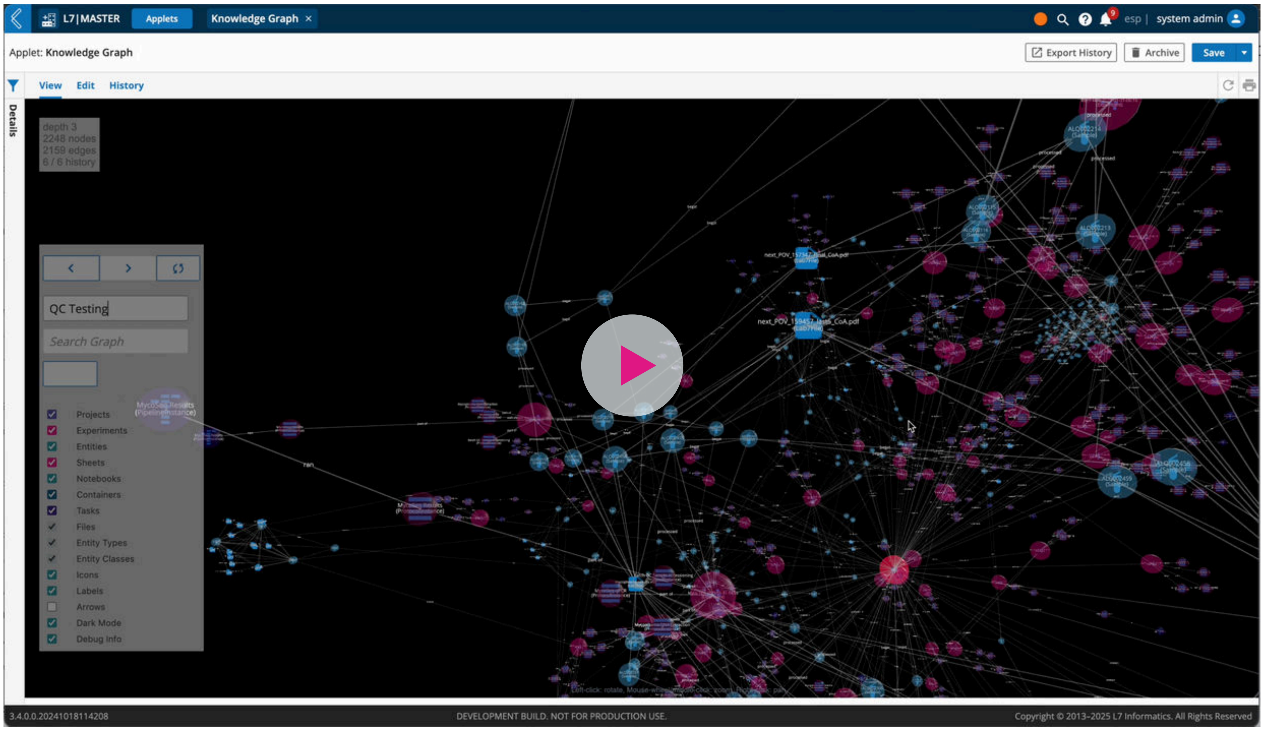
With L7|ESP, you will have full control of your business entities and process model definitions, in an unlocked environment where your data is yours.

Knowledge Graph Generation

L7|ESP contains a graph dataset at its core, enabling the generation of a comprehensive Knowledge Graph (KG) of its data (see Video 1). KGs provide a structured representation of knowledge by detailing entities and their interrelationships. Gen AI models can leverage KGs to achieve better contextual understanding and factual accuracy in generated outputs. Knowledge Graphs in L7|ESP make your structured and unstructured data fully connected, linking every quantitative data point with all its interrelationships (see also Figure 1 above).

More importantly, the data in this graph comes from the different apps in the L7|ESP platform, including L7 LIMS, L7 MES, L7 Notebooks, and L7 Scheduling, which is contextualized and seamlessly integrated following the Reference Models.

This is a key enabler to L7|ESP Sample Provenance and Lot Genealogy features, and the basis for Chain of Identity and Chain of Custody reports.



depth 3
2248 nodes
2159 edges
6 / 6 history

QC Testing

☑ Projects
☑ Experiments
☑ Entities
☑ Sheets
☑ Notebooks
☑ Containers
☑ Tasks
✓ Files
✓ Entity Types
✓ Entity Classes
☑ Icons
☑ Labels
□ Arrows
☑ Dark Mode
☑ Debug Info

Video 1: Knowledge graph of L7|ESP data, harmonized and structured at the point of data capture, which includes contextualized data.

Robust and Normalized Data Model with L7|INTELLIGENCE

L7|INTELLIGENCE offers a robust and normalized data model that simplifies access to data. This streamlined data model ensures that Gen AI models can easily retrieve and process high-quality data, leading to improved training efficiency and inference output quality. To learn more, read the L7|INTELLIGENCE Data Sheet.

Semantic Enrichment via the Integration of Biomedical Ontologies

L7|ESP supports integration with ontologies, allowing different L7|MASTER artifacts to be mapped directly to ontology elements. This integration facilitates semantic enrichment, enabling Gen AI models to understand domain-specific terminology and relationships accurately. For example, in a life sciences application, mapping to biomedical ontologies ensures that the Gen AI model comprehends and generates medically accurate content.

You can read the L7 and SciBite Partnership Enables AI Implementation Through Ontology-backed Data Unification and Harmonization blog post to learn more about how L7|ESP supports the integration with ontologies or watch the “SicBite + L7: Foundations for Effective AI Implementation” webinar which provides some great examples of ontology integration and how they can be applied at the point of data capture.

Process Orchestration Management

L7|ESP’s workflow orchestration engine allows any life science process to be automated, executed, and optimized, leaving automated regulatory-compliant trails along the way while contextualizing process and business data. New child entities generated in a process will have a provenance trail to their parent entities and will be linked to the data captured during process execution in an auditable and secure way.

Gen AI application workflows can leverage workflow orchestration to make the applications GxP compliant, to the extent current LLM technology allows. They can also enable semi-automated, human-in-the-loop supervision and approvals where needed.

Likewise, Agentic AI processes can use this workflow orchestration to execute coordinated agent work. These agents, in turn, can leverage L7|ESP Connectors for their autonomous or semi-autonomous actions, leaving required trails for regulatory compliance, troubleshooting, and, ultimately, accountability.

L7|ESP IS POISED TO ENABLE DISRUPTIVE GEN AI APPLICATIONS IN YOUR COMPANY

Data quality, well-structured data models, and workflow orchestration are pillars of successful Gen AI and Agentic AI applications. L7|ESP offers a comprehensive suite of tools and capabilities that leverage knowledge graphs, robust data models, ontology integration, and workflow orchestration to enhance the performance of Gen AI models and enable the creation of Agentic AI applications. By ensuring high-quality data and structured information, L7|ESP enables enterprise organizations to harness Gen AI's full potential, driving innovation and efficiency in their operations.

> Whether through advanced techniques like RAG and model chaining or through domain-specific applications, L7|ESP stands out as a pivotal solution for optimizing Gen AI performance and enabling the creation of new Agentic AI applications

L7 INFORMATICS

L7INFORMATICS.com

1219 West 6th Street,
Austin, TX 78703 USA
+1 888 461 5227
info@L7informatics.com

L7 Informatics reimagines data intelligence for modern life sciences and healthcare organizations. Beyond simple data management, L7 provides tools that optimize the flow of information between processes and people, unlocking innovation at every stage of the clinical, research, and manufacturing value chains.