

Maximizing your CMC & Manufacturing AI Initiatives

Unlocking the power of data with LLMs

Ken Forman | Lead Product Manager

22 October 2025



IDBS



Vision

Advancing human health by accelerating life-changing therapies



Mission

Helping our customers improve the lives of patients by unlocking the power of data for faster and smarter decisions throughout the BioPharmaceutical lifecycle

IDBS at a glance



35 Years

Experience delivering scientific **Enterprise Lab Informatics** workflows and software via global support teams



75%

Market presence and customer loyalty. 15 of the top 20 biopharma are 10+ year customers¹



Top Awards

Industry-recognized, award-winning products²



Customer Satisfaction

Highest satisfaction and most used scientific lab data platform³



GxP Excellence

Top rated for GxP compliance support & experience⁴



Quality Standards

ISO, SOC & ESG certifications. Platinum pharma supplier status⁵

1. Rank measured by [Market Cap, Q3 2024](#). & 2025 Data on file.

2. [Top Life Sciences Informatics Company in Europe \(Life Sciences Review\)](#), [Best Biologics Digitalisation Award: Data Analytics for Quality Processes \(Asia Pacific Biologics CMO Excellence Awards\)](#), [Top Laboratory Automation Solutions Companies \(pharmatechoutlook.com\)](#), [Best Bioprocessing Supplier Award: Software](#), at the prestigious Asia Pacific Biopharma Excellence Awards (ABEA) 2025.

3. IDBS Market Research study with 856 respondents using ten comparable scientific lab data platforms, 2024-2025.

4. Market research study comparing IDBS to seven other comparable software vendors used in Research & Development, Preclinical CROs and Process Development. Data on file..

5. Awarded platinum pharma supplier status by Qualifyze, the world's leading third-party audit provider for healthcare and pharma companies, 2025.

An industry-leading partner you can trust



Danaher's science and technology leadership puts IDBS' solutions at the forefront of the industry. Together with Danaher's other businesses across Biotechnology, Diagnostics and Life Sciences, we unlock the transformative potential of cutting-edge science and technology to improve billions of lives every day.



Generative AI

Large Language Models

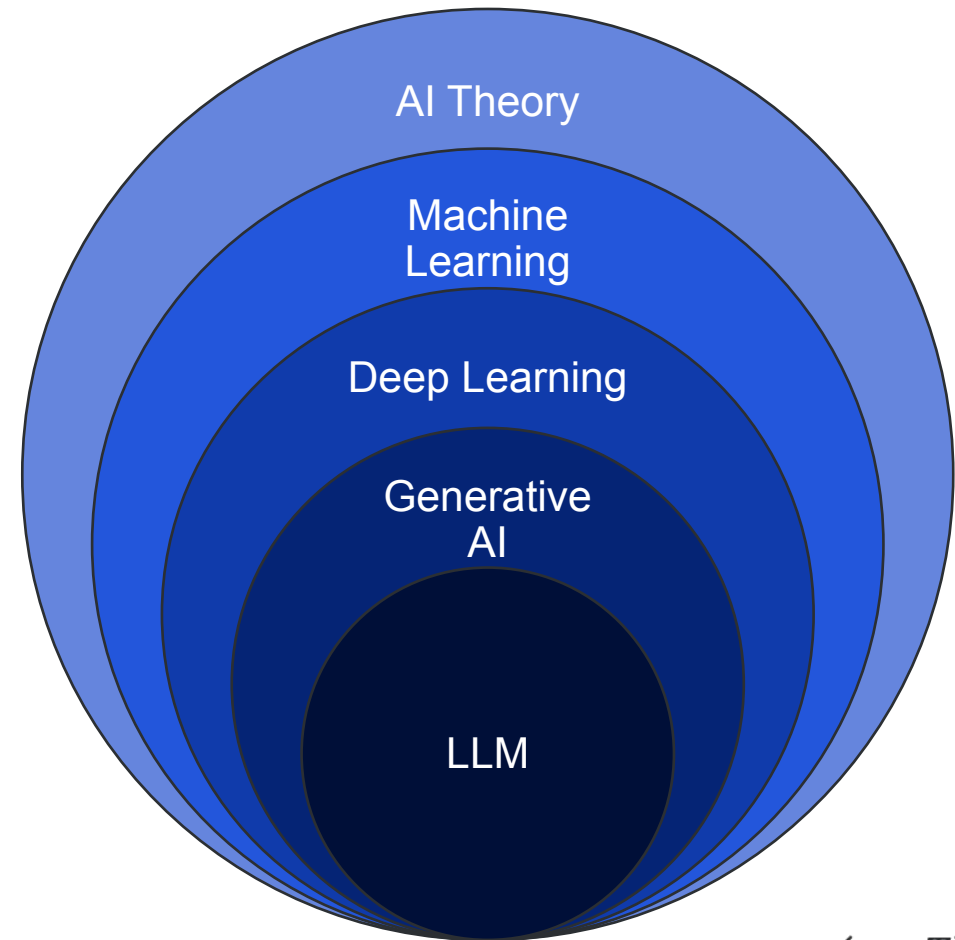
Generative AI is a subset of Deep Learning where the model is trained to create new content given its training data.

Large Language Models are a type of Generative AI which specialize in tasks associated with natural language processing.

Based on the Transformer, which is a deep learning (neural net)

- Text is converted to numerical representations called tokens
- Each token is then converted into a vector via lookup table

You then ask the model to predict the next word (vector result) based on the total vector of your input statement.



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V_6$$

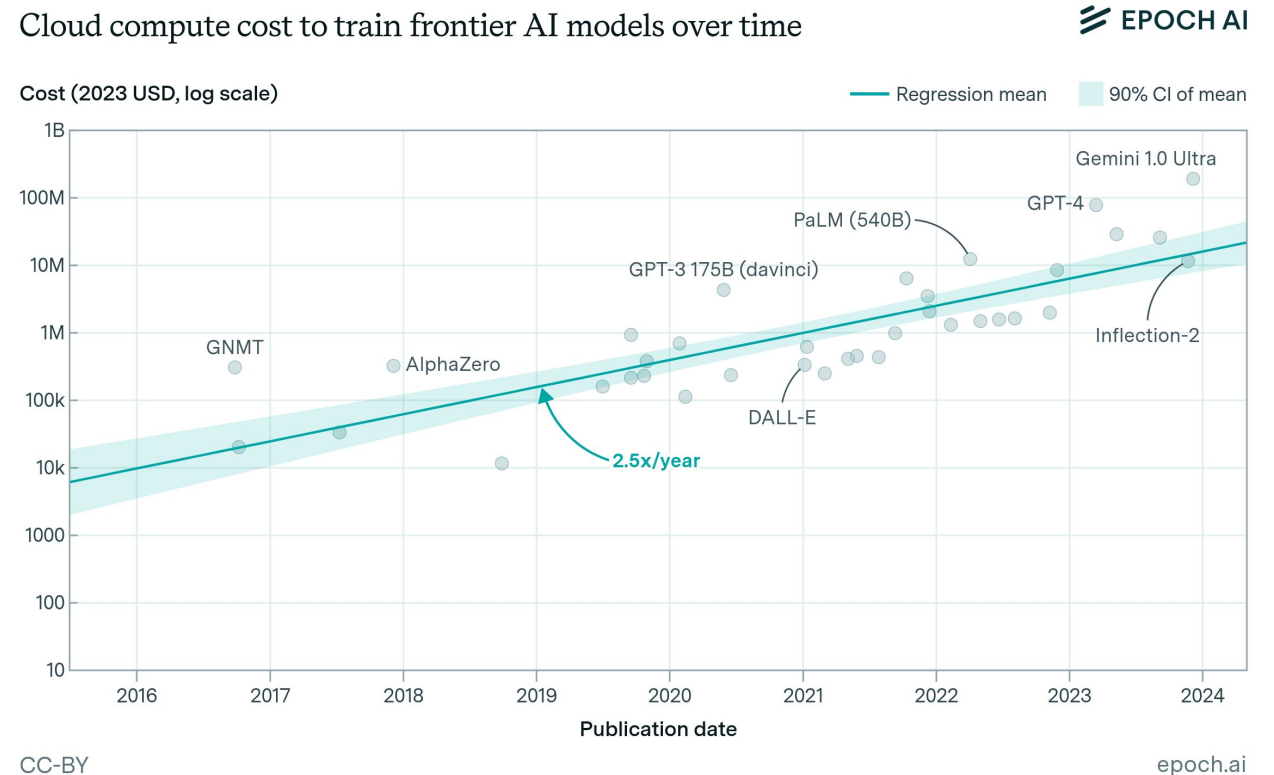
The problem with Large Language Models

Large Language Models are expensive.

- Training of models had increased ~2.5x per year since the invention of the Transformer in 2017
- Primarily due to the increase in size of the data points used to train the model
- For example, GPT-4 has approximately 1.8 trillion parameters, est. to cost ~\$78,000,000 to train
- Note: GPT-5 only has ~635 billion parameters and GPT-5 tells us: *“GPT-5 can’t memorize as much rare or niche knowledge. It relies more on reasoning and retrieval than raw memorization. This means it’s less likely to “know everything” off the top of its head compared to GPT-4.”*

In essence this puts training your own LLM out of reach of most organisations and leaves a growing domain specific problem for biopharma.

Cloud compute cost to train frontier AI models over time



<https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models>

The problem with Large Language Models

Large Language Models are expensive.

- Training of models had increased ~2.5x per year since the invention of the Transformer in 2017
- Primarily due to the increase in size of the data points used to train the model
- For example, GPT-4 has approximately 1.8 trillion parameters, est. to cost ~\$78,000,000 to train
- Note: GPT-5 only has ~635 billion parameters and GPT-5 tells us: *“GPT-5 can’t memorize as much rare or niche knowledge. It relies more on reasoning and retrieval than raw memorization. This means it’s less likely to “know everything” off the top of its head compared to GPT-4.”*

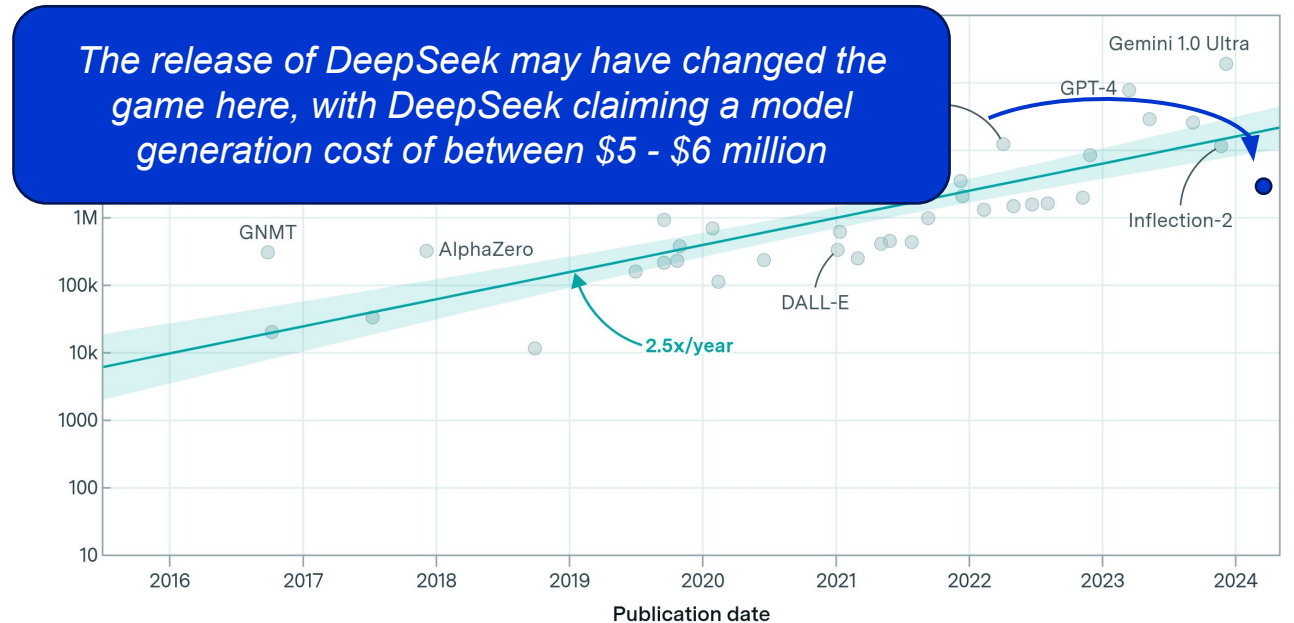
In essence this puts training your own LLM out of reach of most organisations and leaves a growing domain specific problem for biopharma.

Cloud compute cost to train frontier AI models over time

EPOCH AI

Cost (2023 USD, log scale)

— Regression mean
■ 90% CI of mean



CC-BY

epoch.ai

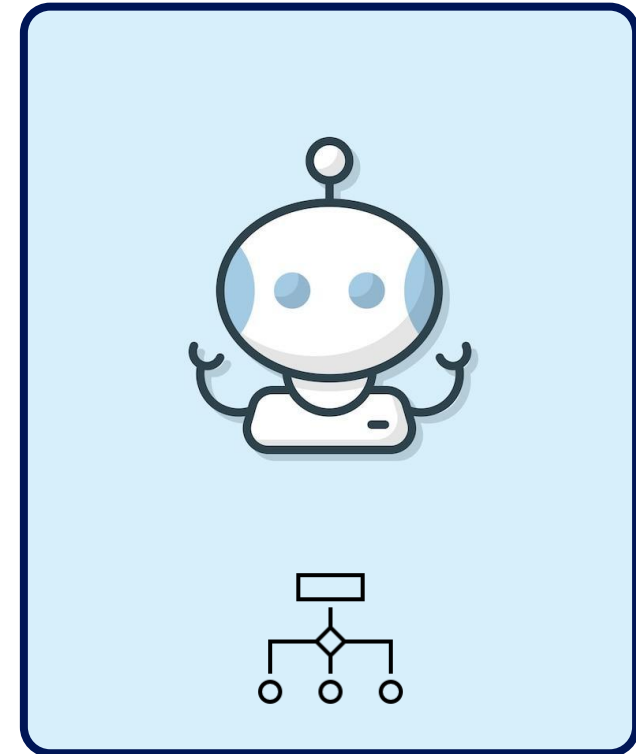
<https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models>

Using Large Language Models in BioPharma

Large Language Models are good but generic, decreasing the specificity and usefulness of responses. If we want to use them in biopharma, we need to make them more domain relevant.

Fortunately, we can through the application of two main methods: retrieval augmented generation (RAG) and Fine Tuning:

- Allows us to adapt LLMs to more specialised tasks
- Uses a smaller dataset than the original model generation
- Generally cheaper to perform
- Allows us to maintain data privacy
- Can leverage Agentic AI for workflow automation, enabling LLMs to act autonomously, orchestrating RAG and fine-tuned models for complex biopharma prompts.



AI Agent

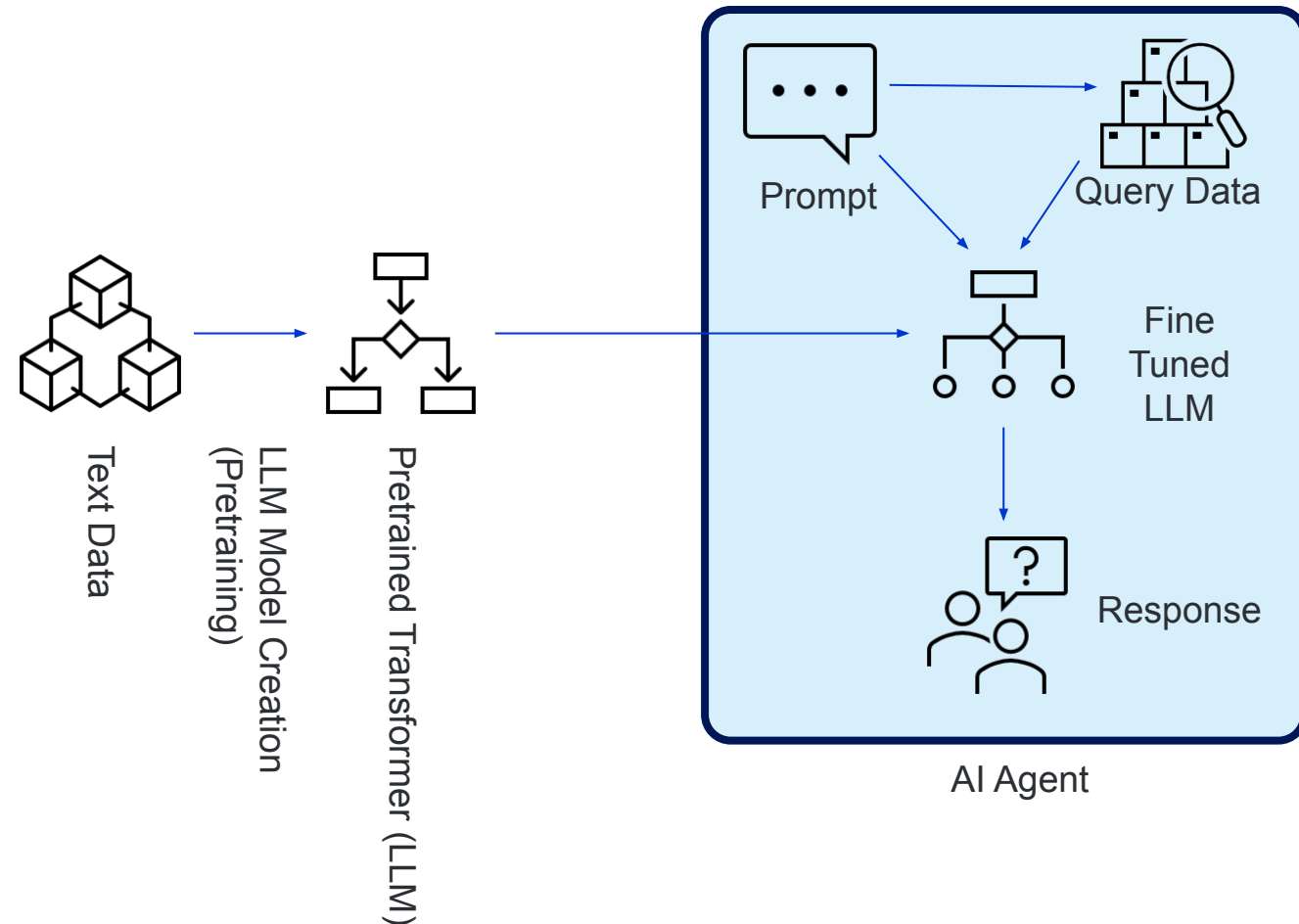
Using Large Language Models in BioPharma: RAG

Retrieval Augmented Generation (RAG) enhances a language model by fetching (retrieving) relevant external information at the time of query. It combines a search/retrieval mechanism with a language model to generate responses using real-time data.

RAG is useful when you need up-to-date, dynamic, or factual information (e.g., customer support)

- **Accesses up-to-date information**
- **No need for model retraining, making it efficient and cost-effective.**
- **Can work with large datasets without increasing model size.**
- **GPT-5 is optimized for RAG!**

- **Slower responses due to the retrieval step.**
- **Dependent on data quality—poor retrieval can lead to poor responses.**



Data quality in a RAG System depends on pre-defined knowledge sources

Curated databases: In a full RAG system, the retrieval step involves searching within pre-defined, curated databases or knowledge bases. These sources are selected based on their relevance and reliability for the specific domain or task.

Biopharma manufacturing examples may include MES, LIMS and historians for process data, QMS and DMS for CAPA's and SOPs.

Semantic search: The system uses advanced semantic search techniques to find the most relevant document segments based on the user's query



AI-generated examples of RAG in Biopharma

1. Real-time quality control adjustments:

- **Scenario:** During a production run, a deviation in product quality is detected.
- **RAG Application:** The system retrieves historical quality control data and inspection reports to identify similar past issues and their resolutions. It then generates recommendations for immediate adjustments to the production parameters to correct the deviation and maintain product quality.

2. Predictive maintenance for equipment:

- **Scenario:** A critical piece of manufacturing equipment shows signs of potential failure.
- **RAG Application:** The system retrieves maintenance logs, sensor data, and historical failure records to predict the likelihood of equipment failure. It generates a maintenance schedule and specific actions to prevent downtime.

3. Optimizing bioprocess parameters:

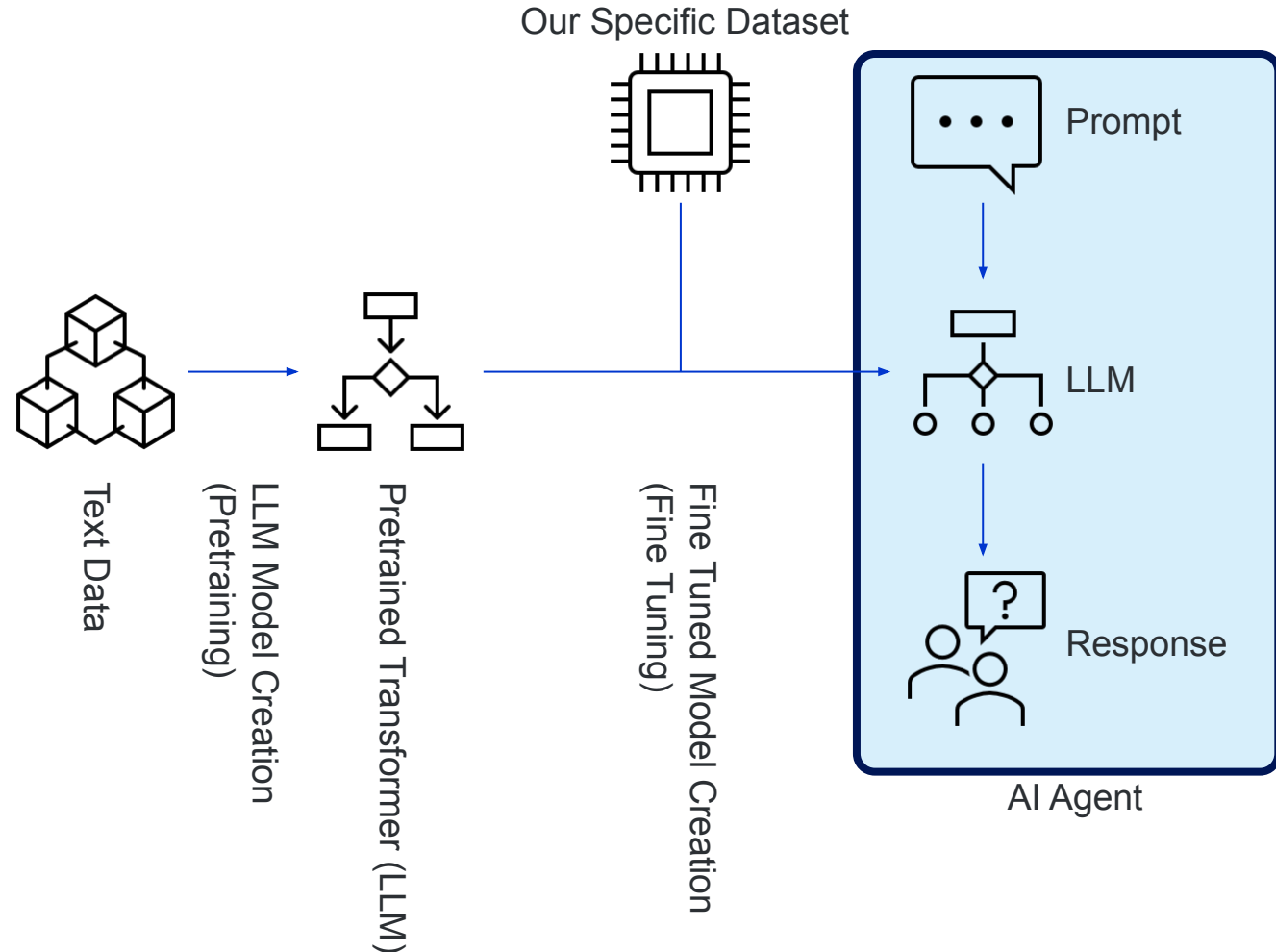
- **Scenario:** Engineers need to optimize the parameters for a new bioprocess to improve yield and efficiency.
- **RAG Application:** The system retrieves relevant research papers, past experiment data, and industry best practices. It generates a set of optimized parameters and experimental conditions tailored to the specific bioprocess.

Using Large Language Models in BioPharma: Fine Tuning

Fine-tuning involves modifying the model's internal weights by training it on a specific dataset. This permanently adjusts the model's knowledge and response style.

Fine Tuning is for when you want personalized, consistent, and structured responses (e.g., brand-specific chatbots, creative writing).

- Produces faster responses (no retrieval step)
- Customizes responses with domain-specific tone, style, or patterns.
- Works well for structured and repetitive tasks.
- **Expensive & time-consuming (requires computing resources).**
- **Requires retraining if new knowledge needs to be added**
- **Dependent on data quality—poor training can lead to poor responses.**



AI-generated examples of Fine Tuning in Biopharma

Optimizing resin reuse in a chromatography skid to reduce costs while maintaining separation efficiency

1. **Initial Setup:** The pre-trained model starts with a resin regeneration protocol based on broad chromatography data.
2. **Fine-Tuning with specific data:** The model is further trained using historical data from the chromatography skid, including resin performance, separation efficiency, and regeneration cycles.
3. **Adjusting regeneration protocol:** during fine-tuning, the model adjusts the resin regeneration protocol to optimize reuse while maintaining separation efficiency. This involves:
 - **Analyzing historical data:** reviewing past chromatography performance data to understand the impact of different regeneration protocols on resin efficiency.
 - **Testing different protocols:** experimenting with various regeneration protocols to find the optimal setting that maximizes resin reuse without compromising separation efficiency.
 - **Validation:** validating the adjusted regeneration protocol with a separate set of data to ensure it performs well under real-world conditions.

Outcome: By fine-tuning the resin regeneration protocol, the chromatography skid achieves efficient resin reuse, reducing costs and maintaining high separation performance

Example of a Fine Tuned LLM

GPT-2

- GPT-2 is a transformer model pretrained on a very large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was trained to guess the next word in sentences.

BioMedLM 2.7B

- BioMedLM 2.7B is a new language model trained exclusively on biomedical abstracts and papers from The Pile. This GPT-style model can achieve strong results on a variety of biomedical NLP tasks, including a new state of the art performance of 50.3% accuracy on the MedQA biomedical question answering task.

```
GPT_FineTune.py > ...
1 # Load model directly
2 from transformers import AutoTokenizer, AutoModelForCausalLM
3 import torch
4
5 prompt = "Summerise the recent research into chromatography?"
6
7 models = ["openai-community/gpt2", "stanford-crfm/BioMedLM"]
8 response = {}
9
10 for m in models:
11     tokenizer = AutoTokenizer.from_pretrained(m)
12     model = AutoModelForCausalLM.from_pretrained(m)
13
14     input_ids = tokenizer(prompt, return_tensors="pt").input_ids
15
16     gen_tokens = model.generate(
17         input_ids,
18         do_sample=True,
19         temperature=0.9,
20         max_length=200,
21     )
22
23     gen_text = tokenizer.batch_decode(gen_tokens)[0]
24
25     response[m + ': '] = gen_text
26
27 print(response)
```

Example of a Fine Tuned LLM

Question: Summarise the recent research into chromatography?

```
from transformers import AutoTokenizer, AutoModelForCausalLM
import torch

prompt = "Summarise the recent research into chromatography?"

model = ["openai-community/gpt2", "stanford-crfm/BioMedLM"]
response = {}

for i, model in enumerate(model):
    tokenizer = AutoTokenizer.from_pretrained(model)
    model = AutoModelForCausalLM.from_pretrained(model)

    input_ids = tokenizer(prompt, return_tensors="pt").input_ids

    gen_tokens = model.generate(
        input_ids,
        do_sample=True,
        temperature=0.9,
        max_length=200,
    )

    gen_text = tokenizer.batch_decode(gen_tokens)[0]

    response[i] = gen_text

print(response)
```

GPT-2

Well, yes! There is a very serious problem with the efficiency of chromatography to produce a good high-quality product that works for everyone. I have seen some people give up their work because they do not have the time to clean up the mess. They are just able to read and absorb chromatographic information.

This is especially true for non-commercial chemical chromatography, where the problem is the waste.

BioMedLM 2.7B

A recent article describes the use of a sensitive high-performance liquid chromatography method for the determination of the amino acid sequences of the alpha2 and beta2 adrenergic receptors (alpha2A-d(5) or alpha2C(*1) adrenoceptors in rabbit, human, and human adrenoceptors.

This study sought to improve the current understanding of the molecular characteristics of the beta adrenoceptors

Data for Generative AI

What is AI-ready data?

Data is structured, contextualized, traceable, and interoperable so AI can reliably consume, interpret, and act on it

FAIR Principles

Ensure data is discoverable, accessible, interoperable and reusable.



Semantic Layer

Provides a consistent understanding of entities through ontologies, NER, metadata, data model for entities and relationships consistency



ALCOA

Ensure data is attributable, legible, contemporaneous, original, and accurate



Governance Layer

Maintains data integrity and compliance through audit trails and lineage



Accurate

Representative

Timel

Consumabl

Sufficient

In Context

Example of a Fine Tuned LLM

If we're going to fine tune an LLM for a specific task, we need to have data available in sufficient size and quality.

Both data volume and data variety contribute to overall quality of your data

Hypothetically, you are working at an up-and-coming biopharma in Process Development. You have a new molecule which is an Immunoglobulin M (IgM) antibodies, you have been asked to design a process description to manufacture the product, but don't know where to start.

So you decide to ask an AI Agent this question:



Example of improving response with a Fine Tuned LLM

If we're going to fine tune an LLM for a specific task, we need to have data available in sufficient size and quality.

Both data volume and data variety contribute to overall quality of your data

Hypothetically, you are working at an up-and-coming biopharma in Process Development. You have a new molecule which is an Immunoglobulin M (IgM) antibodies, you have been asked to design a process description to manufacture the product, but don't know where to start.

So you decide to ask an AI Agent this question:

Question: Create a Process Description with associated CPP and CQAs for an IgM running in our platform process v3

Example response without a Fine Tuned LLM

If we're going to fine tune an LLM for a specific task, we need to have data available in sufficient size and quality.

Both data volume and data variety contribute to overall quality of your data

Hypothetically, you are working at an up-and-coming biopharma in Process Development. You have a new molecule which is an Immunoglobulin M (IgM) antibodies, you have been asked to design a process description to manufacture the product, but don't know where to start.

So you decide to ask an AI Agent this question:

Question: Create a Process Description with associated CPP and CQAs for an IgM running in our platform process v3

GPT-4o: The process begins with the thawing of a high-producing, genetically engineered mammalian cell line (e.g., CHO cells) in a controlled environment. The cells are expanded in shake flasks and transferred to progressively larger bioreactors. The production bioreactor is operated in a fed-batch mode with optimized feeding strategies to maximize cell growth and productivity.

CPPs:

- Temperature control (37°C with slight variation)
- pH (6.8-7.2)
- Dissolved oxygen (DO) levels (30-50%)

CQAs:

- Cell viability (>90%)
- Titer of IgG (>3 g/L)
- Glycosylation profile

Example response with a Fine Tuned LLM

If we're going to fine tune an LLM for a specific task, we need to have data available in sufficient size and quality.

Both data volume and data variety contribute to overall quality of your data

Hypothetically, you are working at an up-and-coming biopharma in Process Development. You have a new molecule which is an Immunoglobulin M (IgM) antibodies, you have been asked to design a process description to manufacture the product, but don't know where to start.

So you decide to ask an AI Agent this question:

(with a pre-trained model)

Question: Create a Process Description with associated CPP and CQAs for an IgM running in our platform process v3

Possible Pretrained Response: Based on the previous 11 examples of an IgM running being performed over the last 4 years at your company. I would use your standard v3 platform process and start the process characterisation with draft specifications of:

Possible CPPs:

- Temperature control (36 – 38°C): Medium Risk to Quality
- pH (6.8-7.2): High Risk to Quality
- Dissolved oxygen (DO) levels (30-50%): High Risk to Quality

Possible CQAs:

- Cell viability (>90%): Low risk to patient safety
- Aggregation (<80%): High risk to patient safety

Data sources for Fine Tuning LLMs

Possible Pretrained Response: Based on the previous 11 examples of an IgM running being performed over the last 4 years at your company. I would use your standard v3 platform process and start the process characterisation with draft specifications of:

Possible CPPs:

- Temperature control (36 – 38°C): Medium Risk to Quality
- pH (6.8-7.2): High Risk to Quality
- Dissolved oxygen (DO) levels (30-50%): High Risk to Quality

Possible CQAs:

- Cell viability (>90%): Low risk to patient safety
- Aggregation (<80%): High risk to patient safety

Historical Experiment Execution Data

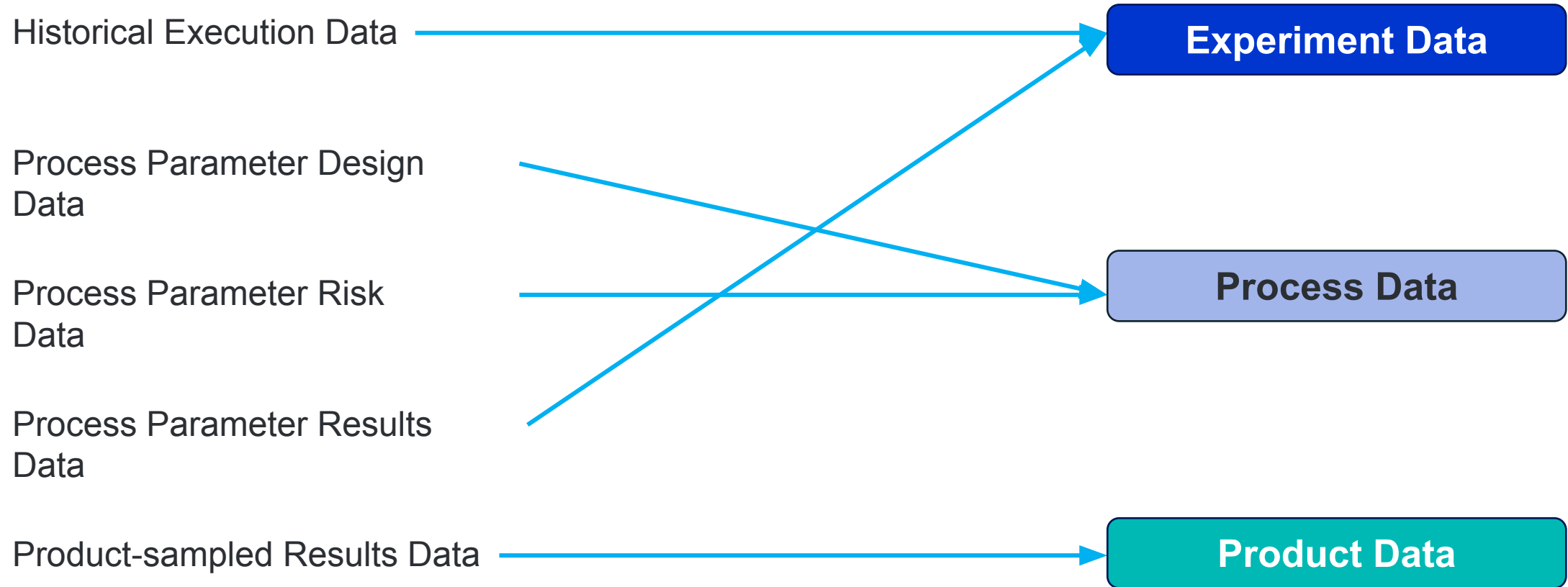
Process Parameter Design Data

Process Parameter Risk Data

Process Parameter Results Data

Product Sample Results Data

Data types for Fine Tuning LLMs



Data systems for Fine Tuning LLMs

Experiment Focused Systems



ELN

Experiment Data



Paper

Process Data

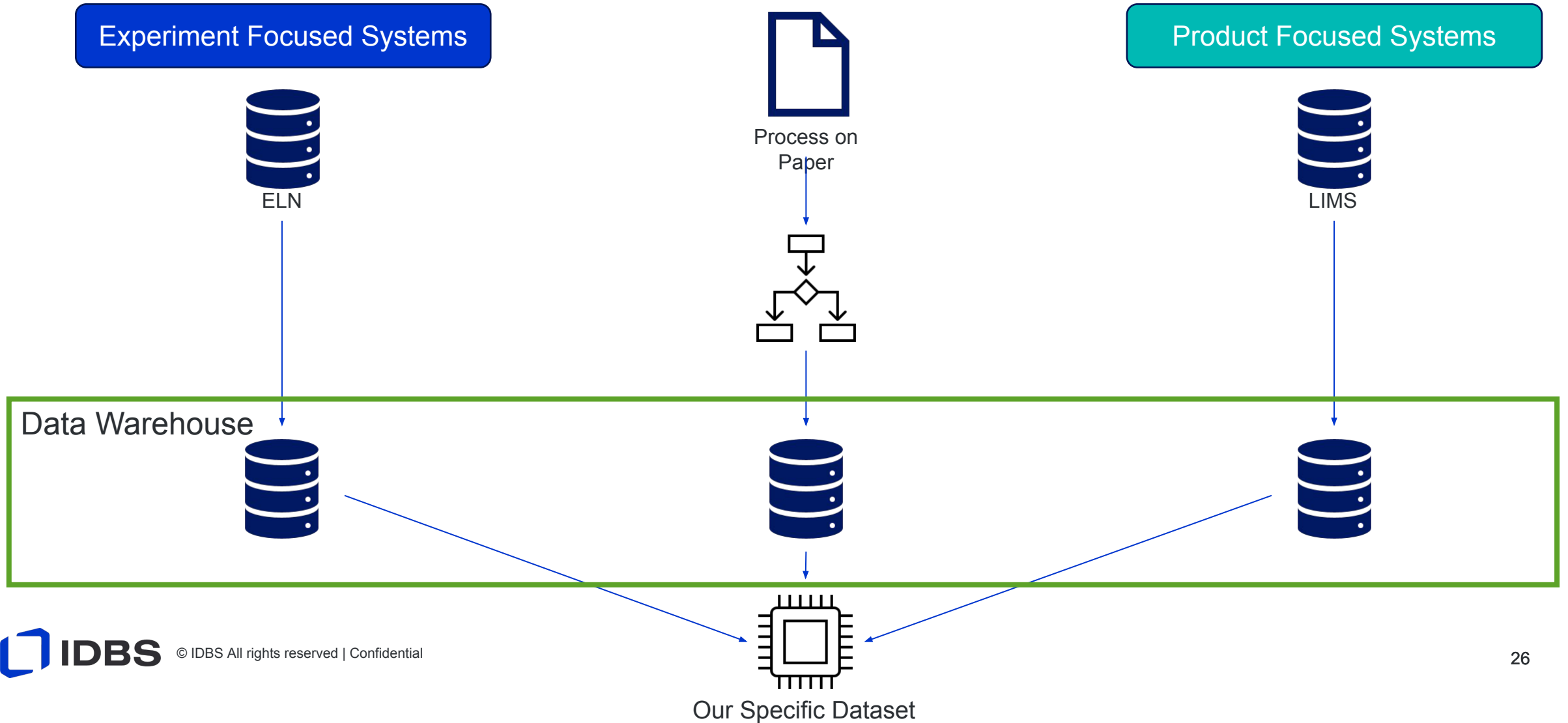
Product Focused Systems



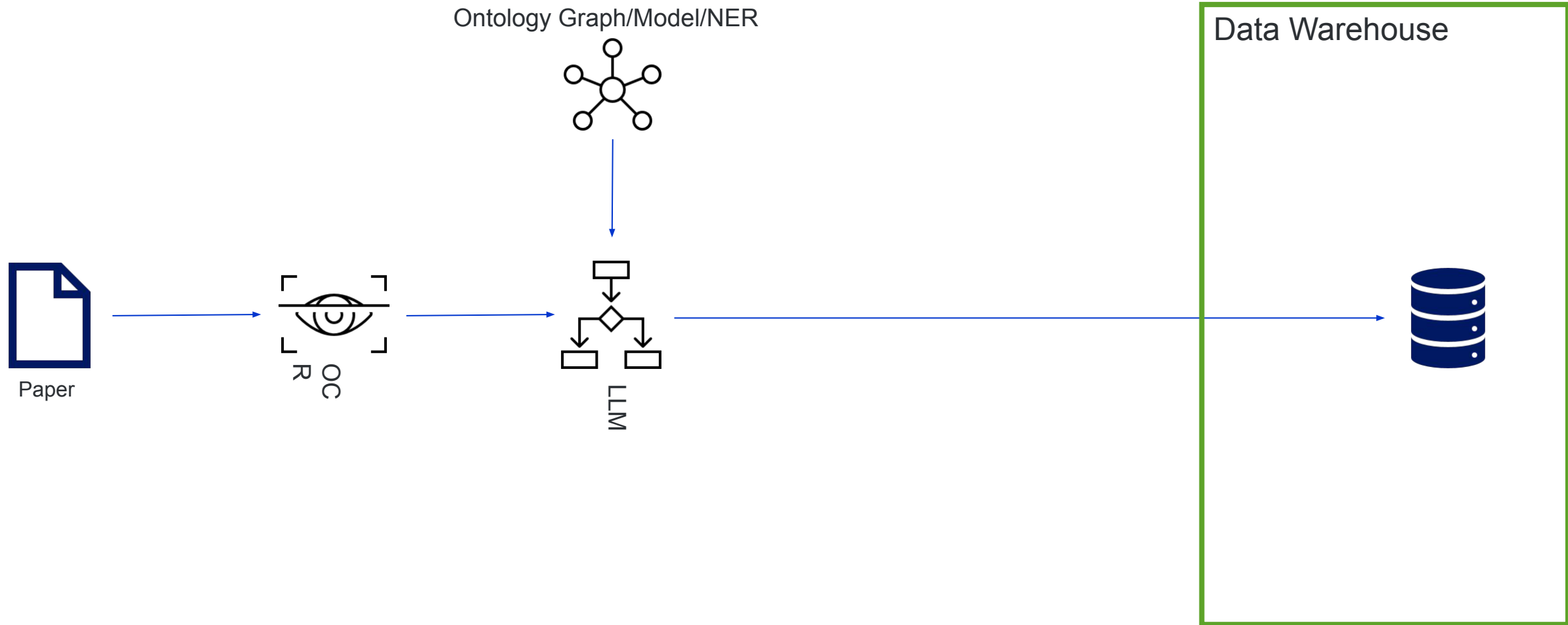
LIMS

Product Data

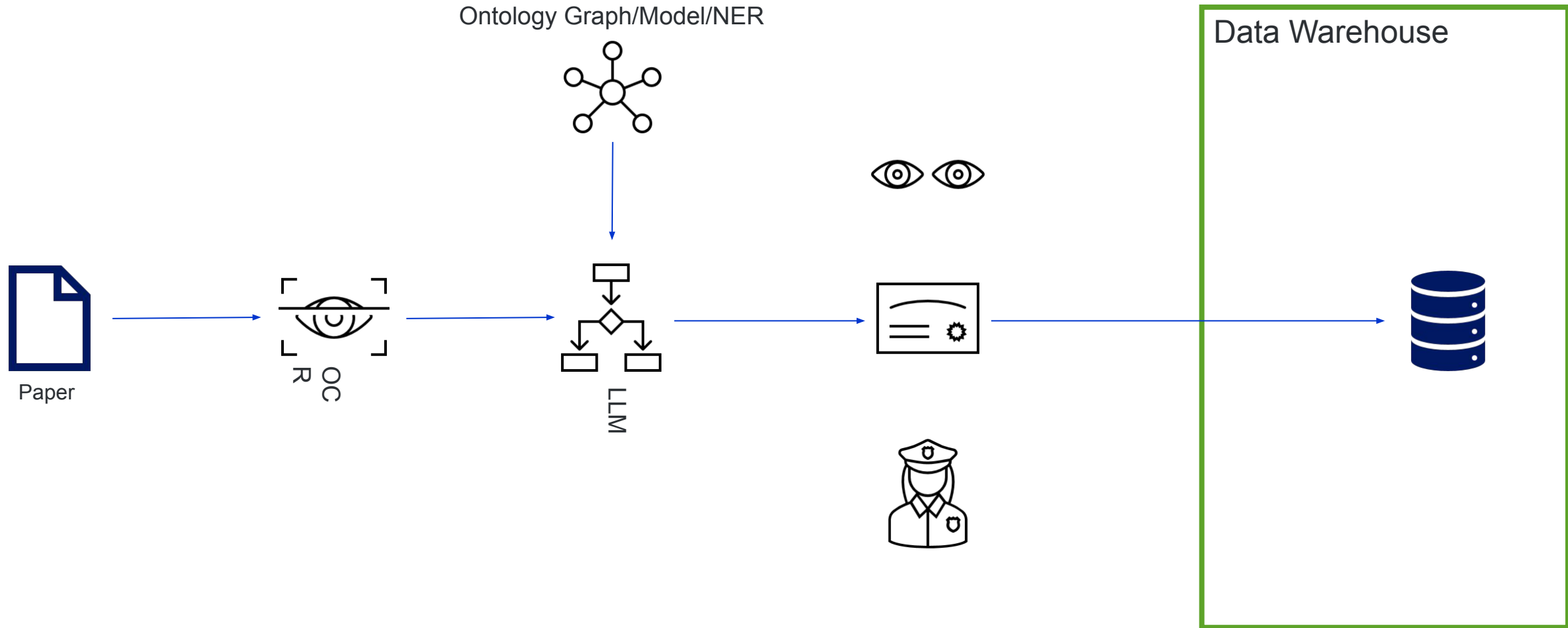
Alignment with a Data Warehouse



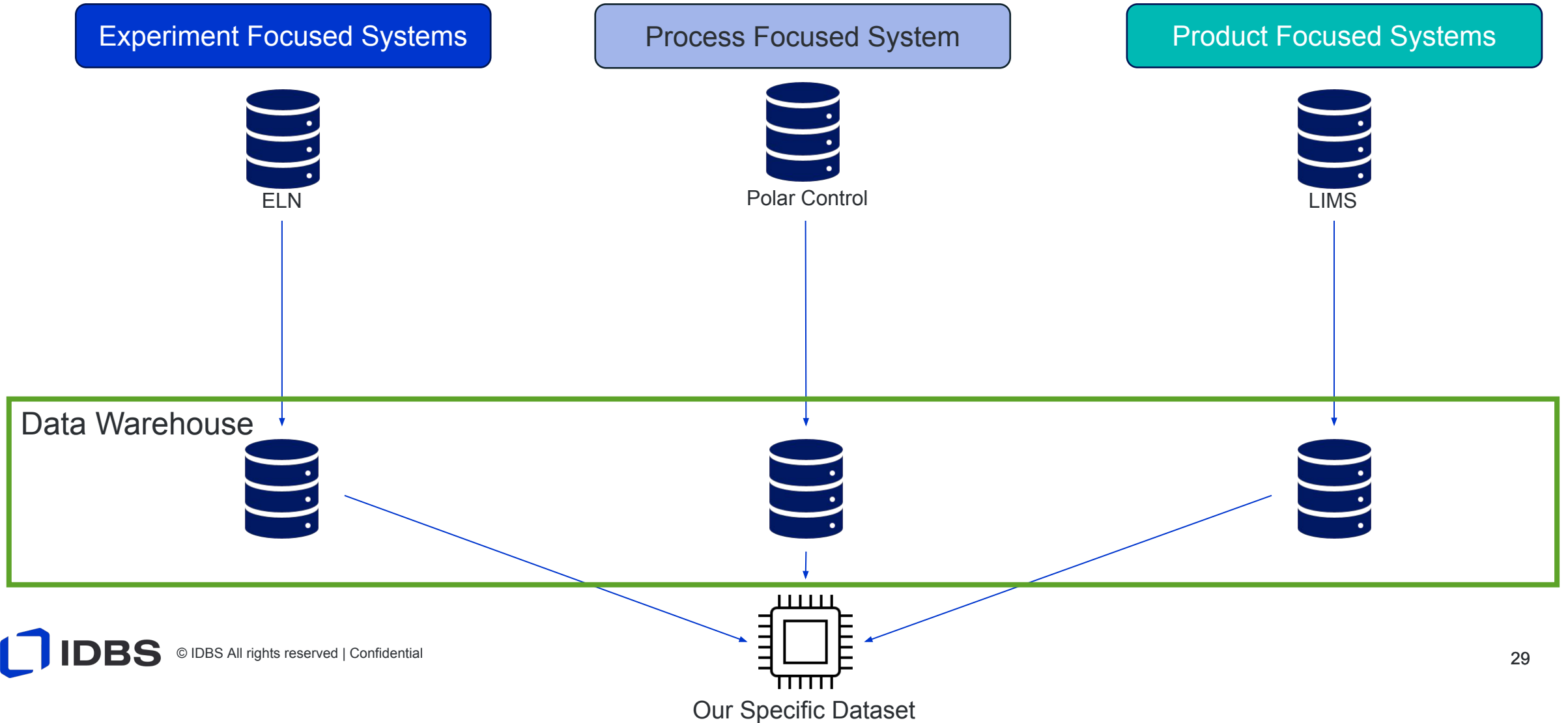
Alignment with a Data Warehouse



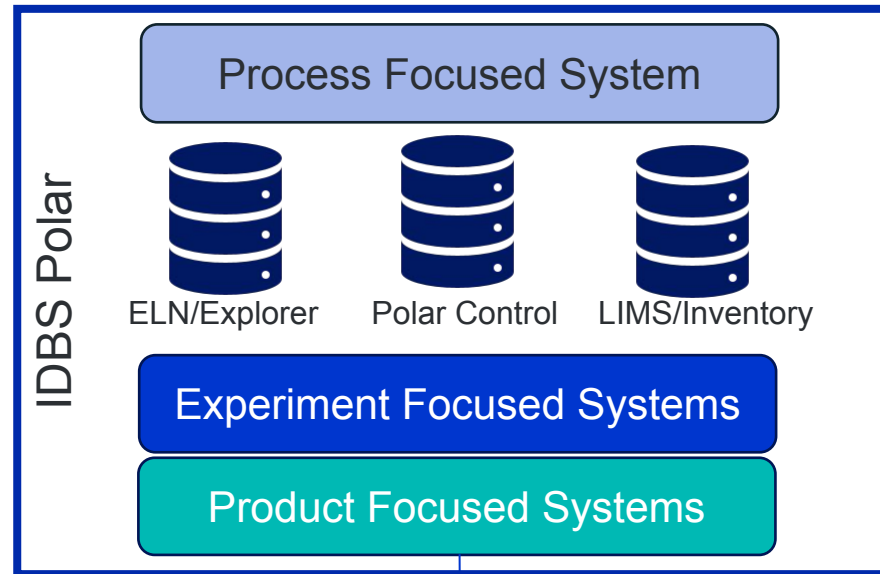
Alignment with a Data Warehouse



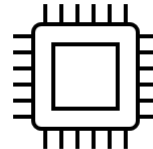
Alignment with a Data Warehouse



Alignment with a Data Warehouse



Data Warehouse



Our Specific Dataset

Structured, contextualized data for Fine Tuned LLMs

Transforming the CMC Landscape

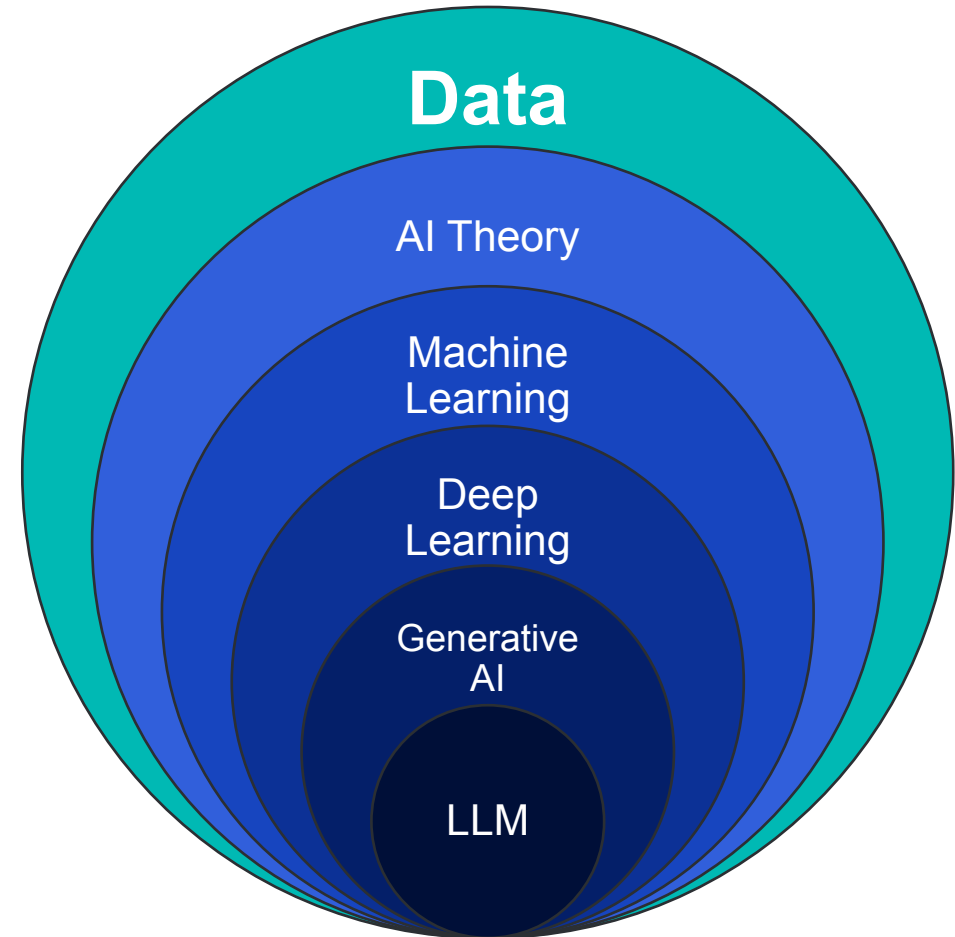
- AI enables smarter, data-driven decision-making in BioPharma
- High-quality, structured data is essential for effective AI adoption

Challenges of Large Language Models (LLMs)

- Expensive to train (e.g., GPT-4 ~\$78M)
- Generic models require fine-tuning for industry-specific tasks

Fine-Tuning for BioPharma Applications

- Adapts LLMs for specialized biopharma needs
- Requires high-quality, diverse data (historical, experimental, process)
- Structured data & warehouse integration are critical



Lively interlude

Question: Show me a glass of wine, full to the brim,
near a clock showing analogue 4:15 in the afternoon.

Lively interlude (Updated)

Question: Show me a glass of wine, full to the brim, near a clock showing analogue 4:15 in the afternoon.





Thank you, any questions?

To learn more about IDBS Polar and other IDBS solutions, please visit the IDBS team at our booth 2-D.

<https://www.idbs.com>