AI Data Curation Market Report

Global Market Segmentation & Forecasts 2025 – 2030

19 March 2026

Document id: IM#1514

Contents

For more information visit https://informationmatters.net

Summary

The transition of artificial intelligence from a experimental novelty to a foundational layer of global industrial operations has necessitated a paradigm shift in data management. As the industry moves past the era of model-centric development, the "curation" layer — encompassing the ingestion, refinement, structuring, and governance of information — has emerged as the primary determinant of model performance and safety. This report provides a comprehensive valuation and strategic forecast of the AI Data Curation sector, structured through a definitive taxonomy of five primary domains.

Our Methodology

The valuation and forecasting in this report are derived from a multi-stage synthesis of primary industrial market reports and secondary data analysis, using the provided 5x5 sector taxonomy as the analytical framework. The research involved mapping each of the 25 sub-categories to their closest industrial equivalents, such as AI Training Datasets, AI Governance, Data Observability, and Knowledge Management Systems. Global market baselines for 2025 were established through a consensus of leading institutional datasets.

Key assumptions underpinning the analysis include:

1. **Currency and Exchange Rates**: All valuations are expressed in US Dollars based on 31 December 2025 spot exchange rates.
2. **Regional Segmentation:** Breakdown by World Bank geographic regions was calculated using 2025 infrastructure and digital adoption benchmarks. Baseline regional shares were weighted by category focus, with infrastructure-heavy segments favoring North America (43.05%) and regulatory-centric segments reflecting higher distribution in Europe (25.0%) and fast-adopting Asian markets (30.0%). Developing regions were sized based on a combination of digital development metrics and venture capital flows.
3. **Growth Modeling**: Forecasts for 2026–2030 were generated by applying sector-specific Compound Annual Growth Rates (CAGR) to 2025 baselines. These rates vary by pillar: Ingestion & Transformation (22.3%), Quality & Governance (36.0%), Knowledge Engineering (43.7%), Training Data (22.9%), and Discovery & Search (14.0%).
4. **Technological Mapping:** Nascent sub-categories such as Agentic Memory and GraphRAG were modeled as subsets of the broader Agentic AI and Knowledge Management markets to ensure accurate capture of the curation layer's total addressable market.

AI Data Curation 5x5 Matrix)

Ingestion & Transformation	Quality, Governance & Trust	Knowledge Engineering & Context	Training Data & Enrichment	Discovery & Search
Unstructured Data Ingestion & Extraction	Data Quality, Governance and Observability	GraphRAG & Structured Context	Data Labeling & Training Infrastructure	Enterprise Search
Unstructured Data Preprocessing	Privacy-First Data Curation	Agentic Memory & Long-Term Context	Synthetic Data Generation	Intelligent Search & AI Relevance
Unstructured Data Management	Trusted RAG & Hallucination Control	Agentic Knowledge Engineering	Data-Centric AI & Quality Control	Real-Time RAG & Search
Real-Time Data Frameworks & Streaming ETL	AI Evaluation & Guardrails	Active Metadata & AI Lineage	Multimodal Curation for Physical AI	Data Framework & Orchestration
Multimodal Data Pipelines	-	Master Data Curation	-	Medical Audio Curation

The 5x5 data curation taxonomy is a structured framework that categorizes the AI Data Curation sector into five primary pillars, each containing up to five distinct sub-categories. This taxonomy maps the entire lifecycle of data as it is prepared, governed, and utilized by artificial intelligence systems.

The five primary pillars are:

1. **Ingestion & Transformation**: This domain focuses on the architecture required to move raw information into AI environments, covering unstructured data extraction, preprocessing, and real-time streaming ETL.

2. **Quality, Governance & Trust:** This pillar serves as the "guardrail" layer, encompassing data observability, privacy-first anonymization, and hallucination control for Retrieval-Augmented Generation (RAG).

3. **Knowledge Engineering & Context**: This section focuses on providing AI with long-term memory and structured reasoning through GraphRAG, agentic memory, and active metadata lineage.

4. **Training Data & Enrichment**: Representing the "engine room" of AI, this category covers traditional data labeling, synthetic data generation, and multimodal curation for physical AI.

5. **Discovery & Search:** The final pillar manages the "last mile" delivery of information via RAG-enabled enterprise search and intelligent discovery platforms.

AI Data Curation Global Market Forecast 2025 - 2030 (USD bn)

	2025	2026	2027	2028	2029	2030	CAGR (%)
North America	32.74	40.54	50.43	63.06	79.29	100.37	25.10%
Asia	23.11	28.62	35.65	44.72	56.42	71.7	25.40%
Europe	18.79	23.31	29.03	36.38	45.82	58.1	25.30%
South America	3.35	4.15	5.17	6.49	8.17	10.35	25.30%
Africa	2.27	2.82	3.52	4.42	5.57	7.08	25.50%
Oceania	1.79	2.22	2.79	3.49	4.43	5.63	25.70%
Global Total	**82.05**	**101.66**	**126.59**	**158.56**	**199.7**	**253.23**	**25.30%**

Ingestion & Transformation: The Architecture of the Zettabyte Era

The Ingestion & Transformation segment represents the critical gateway through which raw, disorganized information enters the artificial intelligence ecosystem. In 2025, the volume of unstructured data is projected to surpass 175 zettabytes, and the inability of legacy systems to process this deluge has created a massive market for specialized ingestion and extraction tools. This segment is characterized by a shift from rigid, batch-based ETL (Extract, Transform, Load) processes toward fluid, real-time, and multimodal data pipelines that can adapt to the shifting requirements of generative AI models.

The primary driver here is the recognition that the majority of enterprise value is trapped in non-tabular formats, such as legal contracts, technical manuals, and sensor logs. Advanced ingestion tools now utilize "intelligent extraction" powered by large language models to preserve the semantic layout of documents, ensuring that structural context is not lost during the conversion to machine-readable formats. This capability is transforming industries like investment banking, where researchers have reported reducing analysis cycles from 40 hours to under 90 minutes by automating the extraction of data from complex financial reports.

Unstructured Data Ingestion & Extraction

The mechanisms of unstructured data ingestion have moved beyond basic Optical Character Recognition (OCR) to include layout-aware transformers. These systems analyze the spatial relationship between text blocks, tables, and images to create high-fidelity digital replicas of physical or PDF assets. The market for these tools is expanding at a CAGR of 22.3% as organizations seek to operationalize their internal document repositories. The demand is highest in North America, where the density of legal and financial services creates a persistent need for high-speed document processing.

Unstructured Data Preprocessing

Preprocessing has evolved into a strategic necessity for reducing downstream compute costs. By deduplicating massive datasets and removing high-entropy "noise" before the data reaches the model training or retrieval phase, preprocessing tools can improve pipeline efficiency by up to 30%. This involves complex tasks such as resolution normalization for multimodal data and the stripping of boilerplate code or web artifacts from training corpora.

Unstructured Data Management

Management tools provide the storage and indexing layer that bridges the gap between raw data lakes and AI-ready datasets. These systems are increasingly adopting "lakehouse" architectures that combine the flexibility of unstructured storage with the governance and searchability of a traditional database. Market leaders like Microsoft and AWS have integrated these management tools into their core cloud offerings, accounting for roughly 35% of the sector's revenue through scalable data lake formation tools.

Real-Time Data Frameworks & Streaming ETL
The advent of "fresh" AI—systems that can reason over information generated only seconds prior—has driven a transition toward streaming ETL. Real-time frameworks allow for the continuous flow of information from live sources like social media feeds, IoT sensors, and financial tickers into the AI's context window. This sub-category is particularly vital for fintech and supply chain applications, where latency is the primary enemy of competitive advantage.

Multimodal Data Pipelines
As models become natively multimodal, the pipelines supporting them must handle text, image, video, and audio simultaneously. The complexity of curating these disparate streams into a synchronized training or retrieval set is immense. These pipelines require advanced orchestration to maintain data lineage and ensure that a change in an image's metadata is reflected across all associated text descriptions.

Ingestion & Transformation Market Forecast 2025 - 2030 (USD bn)

	2025	2026	2027	2028	2029	2030	CAGR (%)
North America	19.25	23.54	28.79	35.21	43.06	52.73	22.30%
Asia	11.18	13.67	16.72	20.45	25.01	30.62	22.30%
Europe	10.37	12.69	15.51	18.98	23.21	28.42	22.30%
South America	1.79	2.19	2.67	3.27	4	4.9	22.30%
Africa	1.23	1.5	1.84	2.25	2.75	3.37	22.30%
Oceania	0.89	1.09	1.34	1.63	2	2.45	22.30%
Total	**44.71**	**54.68**	**66.87**	**81.79**	**100.03**	**122.48**	**22.30%**

Quality, Governance & Trust: The Guardrail Economy

As organizations move AI models from experimental sandboxes to production environments, the focus has shifted from "can it work" to "is it safe." The Quality, Governance & Trust segment is the fastest-growing sub-sector of the curation market, driven by the uncompromising quality standards of high-stakes industries and the entry of significant regulatory frameworks like the EU AI Act. Trust is no longer a philosophical preference; it is a technical requirement for deployment. The mechanism of this growth is the "observability shift"—moving from reactive monitoring to proactive data reliability engineering. Organizations are increasingly aware that the average enterprise lost USD 12.9 million in 2024 due to undetected data errors impacting AI decisions. This has created a robust market for tools that provide real-time visibility into data health and ensure that the inputs to large language models are accurate, unbiased, and compliant with privacy standards.

Data Quality, Governance and Observability

Data observability tools now utilize machine learning to automatically detect "drift" in data distributions. These platforms integrate with cloud data lakes to provide real-time alerts when data quality falls below an established threshold. In the BFSI (Banking, Financial Services, and Insurance) sector, this is a non-negotiable component of the data stack, ensuring that risk assessment models are not making decisions based on stale or corrupted information.

Privacy-First Data Curation

Privacy-first curation automates the anonymization and de-identification of sensitive data. As global privacy regulations like GDPR and CCPA mature, organizations are using these tools to mask Personally Identifiable Information (PII) before it is processed by AI models. This allows enterprises to derive insights from customer data without violating trust or legal mandates, with the "solutions" segment of this market capturing nearly 68% of governance revenue in 2025.

Trusted RAG & Hallucination Control

Retrieval-Augmented Generation (RAG) is the primary method for grounding AI in facts, but the trust of the "retrieved" data remains a bottleneck. Hallucination control tools perform real-time verification of model outputs against the source material, acting as a filter that blocks responses that cannot be corroborated by the underlying knowledge base. This sub-category is seeing exponential growth in healthcare, where the cost of a factual error can be life-threatening.

AI Evaluation & Guardrails

Evaluation frameworks allow organizations to benchmark models against safety, ethics, and performance standards. Guardrails provide a dynamic protection layer that monitors the "conversation" between a user and an AI, blocking harmful content or attempts to extract sensitive information. Gartner projects that by 2030, AI regulations will quadruple, forcing 75% of the world's economies to spend heavily on these governance technologies.

Quality, Goverance & Trust Market Forecast 2025 - 2030 (USD bn)

	2025	2026	2027	2028	2029	2030	CAGR (%)
North America	1.79	2.43	3.31	4.5	6.12	8.33	36.00%
Asia	1.69	2.3	3.13	4.26	5.79	7.87	36.00%
Europe	1.41	1.92	2.61	3.55	4.82	6.56	36.00%
South America	0.34	0.46	0.63	0.85	1.16	1.57	36.00%
Africa	0.23	0.31	0.42	0.57	0.77	1.05	36.00%
Oceania	0.18	0.25	0.34	0.46	0.63	0.86	36.00%
Total	**5.64**	**7.67**	**10.43**	**14.19**	**19.3**	**26.24**	**36.00%**

Knowledge Engineering & Context: The Foundation of Agentic AI

The evolution of AI from simple chatbots to autonomous agents has revitalized the field of knowledge engineering. While basic Large Language Models (LLMs) are impressive in their reasoning, they are often "memoryless" or limited by a finite context window. The Knowledge Engineering & Context pillar focuses on providing AI with the long-term memory and structured reasoning it needs to function as an effective employee or digital assistant.

The primary technological shift in this sector is the move from Vector-based RAG to GraphRAG. While vectors are excellent at finding "similar" documents, they struggle with "relational" reasoning (e.g., "how is this project connected to our budget in 2023?"). GraphRAG utilizes knowledge graphs to map the actual relationships between entities, allowing an AI agent to navigate complex organizational hierarchies and interconnected data points.

GraphRAG & Structured Context

GraphRAG is the "gold standard" for enterprise context. By structuring information as a graph of nodes and edges, enterprises can ensure that AI agents understand the provenance and relationship of every fact. This sub-category is seeing high adoption in life sciences, where researchers use knowledge graphs to automate complex biomedical workflows and drug discovery.

Agentic Memory & Long-Term Context

AI agents require the ability to maintain "state" over long interactions. Agentic memory systems provide the infrastructure for an agent to remember a user's preferences, past tasks, and the current status of ongoing projects across different sessions. This is critical for the "15% of day-to-day work decisions" that Gartner expects to be made autonomously by 2028.

Agentic Knowledge Engineering

This category involves using AI to build and maintain the knowledge base itself. Instead of human librarians manually tagging documents, "curator agents" crawl through enterprise data, identify key concepts, and structure them into the knowledge graph. This automation is essential for keeping knowledge bases fresh in a world where data is generated faster than humans can read it.

Active Metadata & AI Lineage

Active metadata uses machine learning to dynamically tag data as it is used, while AI lineage provides the "audit trail" that shows exactly how a piece of data was transformed before it reached the model. This is particularly important for compliance with the EU AI Act's transparency requirements, ensuring that every AI decision can be traced back to its underlying data inputs.

Master Data Curation

Master data curation focuses on creating the "golden record"—the single source of truth for critical entities like "Customer" or "Product." In an AI-driven organization, having multiple conflicting records for the same customer leads to catastrophic errors in personalized agents. MDM for AI ensures that every agent in the organization is working from the same foundation.

Knowledge Engineering & Context Market Forecast 2025 - 2030 (USD bn)

	2025	2026	2027	2028	2029	2030	CAGR (%)
North America	3.73	5.37	7.71	11.08	15.92	22.88	43.70%
Asia	2.98	4.28	6.14	8.83	12.69	18.24	43.70%
Europe	2.11	3.04	4.36	6.27	9.01	12.94	43.70%
South America	0.34	0.48	0.69	1	1.43	2.06	43.70%
Africa	0.24	0.35	0.5	0.71	1.02	1.47	43.70%
Oceania	0.2	0.29	0.42	0.6	0.86	1.24	43.70%
Total	**9.6**	**13.8**	**19.82**	**28.49**	**40.94**	**58.83**	**43.70%**

Training Data & Enrichment: The Synthetic Data Revolution

The Training Data & Enrichment segment is the engine room of AI model development. While the "big models" have already been trained on the open internet, the future of AI lies in specialized models trained on high-quality, domain-specific data. The most significant trend in this segment is the "Synthetic Pivot." Analysts estimate that by 2025, real-world data will be exhausted or legally restricted for many use cases, making synthetically generated datasets a primary propellant for the market.

The mechanism here is the "flywheel of enrichment": real-world data is used to seed a generative model, which then creates millions of high-quality synthetic examples, which are then used to train a more robust specialized model. This is particularly effective in autonomous driving, where "edge cases"—like a car driving through a blizzard at night—are too dangerous to collect in the real world but can be synthesized with perfect accuracy.

Data Labeling & Training Infrastructure

While manual labeling still accounts for a significant share of the market, it is being transformed by "human-in-the-loop" (HITL) platforms. These systems use AI to provide the first pass of annotation, which human subject matter experts then verify. This reduces costs by 30-40% while maintaining the "pixel-level" accuracy required for medical imaging and defense applications.

Synthetic Data Generation

Synthetic data is the creation of artificial datasets that possess the same mathematical properties as real data without containing any real PII. This sub-category is growing at a 46.4% CAGR as organizations seek to bypass the "data drought" and comply with privacy regulations. Leading firms like Mostly AI and Gretel are specializing in tabular data synthesis for the banking and healthcare sectors.

Data-Centric AI & Quality Control

Data-Centric AI is a movement that prioritizes the health of the dataset over the complexity of the model. Tools in this sub-category identify "poisoned" data, detect bias, and ensure that the training set is representative. This is critical for preventing "model drift," where an AI's performance degrades over time because the real-world data it encounters no longer matches its training set.

Multimodal Curation for Physical AI

Physical AI—robotics, drones, and autonomous vehicles—requires multimodal curation that fuses camera feeds, LiDAR, and IMU data. This sub-category provides the infrastructure to synchronize these sensors, allowing a robot to learn how to navigate a warehouse or perform delicate surgery. The demand for "scene simulation" is a key driver here, as robots must be trained in digital twins before they touch the real world.

Training Data & Enrichment Market Forecast 2025 - 2030 (USD bn)

	2025	2026	2027	2028	2029	2030	CAGR (%)
North America	1.25	1.54	1.89	2.32	2.85	3.5	22.90%
Asia	1.15	1.41	1.73	2.13	2.62	3.22	22.90%
Europe	0.83	1.02	1.26	1.55	1.9	2.34	22.90%
South America	0.14	0.18	0.22	0.27	0.33	0.4	22.90%
Africa	0.11	0.13	0.16	0.2	0.25	0.3	22.90%
Oceania	0.11	0.13	0.16	0.2	0.25	0.3	22.90%
Total	**3.59**	**4.41**	**5.42**	**6.66**	**8.19**	**10.07**	**22.90%**

Discovery & Search: The Interface of Information

The final pillar of the curation taxonomy is Discovery & Search. This category represents the "last mile" of data curation, where information is delivered to the user. In 2025, the market is undergoing a structural shift from keyword-based search to "answer-based" discovery. Half of all consumers now intentionally seek out AI-powered search engines, and this is expected to impact over $750 billion in consumer spend by 2028.

The mechanism for this growth is the "Search-as-Service" (SaaS) model. Instead of maintaining complex internal search indexes, enterprises are subscribing to AI discovery platforms that can crawl their internal silos and provide natural language answers to employee questions. This is particularly transformative in the retail sector, where "intelligent search" now accounts for 42% of the market share, enabling shoppers to find products through conversational queries rather than rigid filters.

Enterprise Search

Enterprise search has moved from "finding a document" to "finding an answer." RAG-enabled platforms allow employees to query their entire company history and receive a summarized response with citations. This improves productivity significantly, as the average employee in a non-AI-centric organization still recreates up to 80% of lost documents because they cannot find them.

Intelligent Search & AI Relevance

Intelligent search systems use behavioral data to personalize search results in real-time. These engines learn what a user is looking for based on their role, past queries, and current project context. In the e-commerce sector, this has become the primary battleground for customer loyalty, with 50% of consumers already using AI-powered summaries for buying decisions.

Real-Time RAG & Search

Real-time RAG ensures that search results are always "fresh." This is vital for customer service, where an agent needs the most recent policy update, or for financial news, where a five-minute delay can be the difference between profit and loss. The global market for these real-time search platforms is projected to grow at a 14% CAGR.

Data Framework & Orchestration

Search orchestration provides the "plumbing" that keeps search indexes in sync with underlying databases. As organizations move to multi-cloud environments, these frameworks ensure that a search query can span across AWS, Azure, and on-premise servers without the user seeing the complexity.

Medical Audio Curation
Specialized for the healthcare vertical, this sub-category involves the ingestion and search of clinical audio. By transcribing and indexing doctor-patient interactions, these tools allow

Discovery & Search Market Forecast 2025 - 2030 (USD bn)

	2025	2026	2027	2028	2029	2030	CAGR (%)
North America	6.72	7.66	8.73	9.95	11.34	12.93	14.00%
Asia	6.11	6.96	7.93	9.05	10.31	11.75	14.00%
Europe	4.07	4.64	5.29	6.03	6.88	7.84	14.00%
South America	0.74	0.84	0.96	1.1	1.25	1.42	14.00%
Africa	0.46	0.53	0.6	0.69	0.78	0.89	14.00%
Oceania	0.41	0.46	0.53	0.6	0.69	0.78	14.00%
Total	**18.5**	**21.09**	**24.04**	**27.41**	**31.25**	**35.62**	**14.00%**

Drivers of Growth in the AI Data Curation Sector

The expansion of the AI Data Curation sector is not merely a byproduct of the broader AI boom; it is a structural response to the fundamental limitations of first-generation generative AI. As the industry moves from the "wow phase" to the "work phase," the focus has shifted to the integrity and utility of the underlying information. This transition is propelled by a synergy of technical, economic, and social drivers that have made data curation the most strategic layer of the modern technology stack.

Technical Drivers: Multimodality, Context, and the Synthetic Pivot

Technically, the curation market is being driven by the "Multimodal Imperative." The first generation of AI was largely text-based, but 2025 has seen the rise of native multimodal models that can reason across sight, sound, and language simultaneously. This has created an unprecedented technical challenge: how does one "curate" a dataset that includes synchronized video, audio transcripts, and sensor logs? The creation of "Multimodal Data Pipelines" is a direct technical response to this complexity.

Furthermore, the "Context Window Wars" have reached a stalemate. While model providers can now process millions of tokens, doing so is computationally expensive and often leads to the "lost in the middle" phenomenon where the AI ignores the middle of a large prompt. The technical solution has been "Knowledge Engineering" — using GraphRAG and Agentic Memory to select only the most relevant context to feed the model. This is the technical mechanism that allows an AI to act as a long-term "productivity partner" rather than a one-off query engine.

Finally, the technical "Synthetic Pivot" is perhaps the most transformative driver. As the internet's supply of high-quality human data is exhausted, and as the "poisoning" of the web by AI-generated content makes scraping more difficult, organizations are turning to synthetic data generation. This shift from "collecting" data to "generating" curated data is a technical evolution that allows for the creation of perfectly labeled datasets that are free from PII and bias, effectively solving the "data drought" for specialized industries.

Economic Drivers: ROI, Operational Efficiency, and the Cost of Error

Economically, the curation sector is driven by a massive shift in corporate spending from experimental "AI labs" to production-ready "AI operations." The primary economic motivator is the demonstrable ROI of agentic workflows. Google Cloud's 2025 research found that 74% of organizations achieved a return on investment in the first year of AI agent deployment. These economic gains are most visible in customer service—where curated information reduces contact time by 120 seconds—and in security operations, where AI agents have led to a 70%

reduction in breach risk.

However, the "negative ROI" of poor data is an equally powerful economic driver. The average enterprise lost USD 12.9 million in 2024 because of "dirty" data that led to incorrect AI outputs. In sectors like healthcare or finance, a single hallucinated fact can lead to millions in regulatory fines or lawsuits. This has transformed "Quality, Governance & Trust" from a defensive compliance cost into an offensive economic strategy. Organizations that invest in curation are essentially buying "insurance" for their AI investments, ensuring that the millions spent on GPUs and model licenses are not wasted on a system that the organization cannot trust.

Social Drivers: The Privacy Paradox and the Digital Divide

Socially, the curation market is being shaped by the "Privacy Paradox." Users want the hyper-personalized experiences that AI can provide, but they are increasingly distrustful of how their data is being used. This has led to a global movement toward "Privacy-First Data Curation". The EU AI Act is the most prominent social driver in this regard, codifying the public's demand for transparency, safety, and the right to an explanation for AI-driven decisions. Data curation is the only mechanism that allows an organization to satisfy these social demands while still delivering high-performance AI.

Moreover, a significant social driver is the widening "Digital Divide" between the Global North and Global South. While North America and Europe lead in infrastructure spend, regions like Asia and Africa are seeing the fastest growth in AI adoption among the working-age population. In the UAE, for instance, 64% of the population uses AI daily, significantly higher than the US usage rate of 28.3%. This "leapfrogging" behavior in emerging economies is driving the demand for localized and multilingual curation—ensuring that AI models work as effectively in Arabic or Hindi as they do in English. This social need for "Sovereign AI" is forcing governments to invest in domestic curation pipelines that reflect their own cultural and linguistic nuances.

AI Data Curation - Companies to Watch

Ingestion & Transformation	Companies
Unstructured Data Ingestion & Extraction	Reducto Instabase Heptabase RAGFlow
Unstructured Data Preprocessing	Unstructured.io
Unstructured Data Management	Komprise
Real-Time Data Frameworks & Streaming ETL	Qlik (Talend) Pathway Airia Estuary
Multimodal Data Pipelines	Graphlit

Quality, Governance & Trust	Companies
Data Quality, Governance and Observability	Galileo Alation Encord Arize AI Monte Carlo
Privacy-First Data Curation	OneTrust Ketch BigID Skyflow Privacera
Trusted RAG & Hallucination Control	Vectara
AI Evaluation & Guardrails	Patronus AI

AI Data Curation - Companies to Watch (contd..)

Knowledge Engineering & Context	Companies
GraphRAG & Structured Context	Nebula Graph FalkorDB Neo4j
Agentic Memory & Long-Term Context	Cognee Mem0 Zep
Agentic Knowledge Engineering	WhyHow
Active Metadata & AI Lineage	Informatica Atlan Databricks Collibra
Master Data Curation	Tamr

Training Data & Enrichment	Companies
Data Labeling & Training Infrastructure	Scale AI Snorkel AI
Synthetic Data Generation	Mostly AI Ydata Fabric Tonic.ai Gretel.ai K2view
Data-Centric AI & Quality Control	Cleanlab
Multimodal Curation for Physical AI	SuperAnnotate

AI Data Curation - Companies to Watch (contd..)

Discovery & Search	Companies
Enterprise Search	Glean
Intelligent Search & AI Relevance	Coveo
Real-Time RAG & Search	Nuclia
Data Framework & Orchestration	Llamaindex
Medical Audio Curation	Abridge

information matters

Strategic Analysis of the Agentic AI Revolution

About Us

Our core focus is exploring the “why” and “what’s next” in the dynamic Agentic AI space. What are the market opportunities? The impact of Agentic AI on industry sectors. Market forecasts and strategic analysis.

Find out more about Information Matters:

https://informationmatters.net

Email: info@informationmatters.net