AI Training Dataset Market by Dataset Creation (Data Collection, Data Annotation, Synthetic Data Generation), Dataset Selling (Off-the-Shelf Datasets, Dataset Marketplaces), Data Modality (Text, Image, Video, Audio, Multimodal) - Global Forecast to 2029

Market Report | 2024-10-24 | 447 pages | MarketsandMarkets

AVAILABLE LICENSES:

- Single User $4950.00
- Multi User $6650.00
- Corporate License $8150.00
- Enterprise Site License $10000.00

Report description:

The market for AI training datasets is expected to increase from USD 2.82 billion in 2024 to USD 9.58 billion in 2029, experiencing a compound annual growth rate (CAGR) of 27.7% from 2024 to 2029. The demand for AI training datasets is rapidly increasing as various sectors look for more machine learning and AI uses. A key factor driving the growth of the market is the increasing demand for top-notch, varied data collections to properly train AI models, especially in industries such as healthcare, finance, and autonomous vehicles. However, concerns regarding data privacy and compliance with regulations continue to pose a major barrier that could hinder data collection and restrict access to personal data. Businesses encounter difficulties in obtaining and controlling data that comply with performance and regulation requirements, while also harmonizing innovation and ethical factors.
"By offering, dataset creation segment is expected to register the fastest market growth rate during the forecast period."
The dataset creation segment is expected to have the quickest increase in the market in the forecast period, due to the growing need for top-notch data in different industries. Businesses are realizing the significance of making decisions based on data and are therefore making substantial investments in developing thorough and precise sets of data. This part takes advantage of AI and ML progress, which simplify data collection and processing, enabling businesses to create datasets more quickly and on a larger scale. Additionally, the rapid growth of this sector is fueled by the increasing number of IoT devices, and the growing amount of data produced from digital interactions. Companies are prioritizing the creation of large data sets to conduct predictive analysis, comprehend customer actions, and devise tailored marketing tactics to improve their results. Rules like GDPR and CCPA have prompted businesses to focus on ethical ways of collecting data, creating a demand for customized datasets that abide by the regulations. Companies require tailored data sets to meet specific business requirements in order to stay competitive in their

respective industries and experience market growth.
"By dataset selling, Off-the-Shelf (OTS) datasets segment is expected to have the largest market share during the forecast period."
The OTS datasets are expected to lead the dataset selling segment in market because of their inexpensive price, easy access, and immediate suitability for various uses. Companies are opting for pre-made datasets more often as they save time on data collection and preparation, enabling a swift adoption of data-driven strategies. The rising demand for data analysis in different sectors such as healthcare, finance, and marketing are pushing this trend further, as companies seek to leverage existing data for improved decision-making and obtaining valuable insights. In addition, the rise of artificial intelligence and machine learning technologies has raised the demand for top-notch data to train models, resulting in a heavier reliance on pre-made datasets. The use of ready-made datasets is expected to rise steadily in the upcoming years as businesses prioritize adaptability and remaining competitive.
"By annotation type, synthetic datasets segment is expected to register the fastest market growth rate during the forecast period."
Throughout the predicted period, the synthetic datasets segment in the AI training dataset market is expected to experience the most significant increase in growth rate. Synthetic datasets generate abundant data simulating real-world scenarios, solving problems of insufficient data and privacy issues associated with authentic datasets. Customizing synthetic data to suit particular purposes increases its attractiveness, since it can be tailored to fulfill the diverse demands of artificial intelligence models across different industries. Progress in developing models and simulation techniques enhances the accuracy and authenticity of synthetic data, ultimately boosting its efficacy in training machine learning algorithms. The demand for robust and flexible datasets is projected to increase as companies focus on improving their AI capabilities, underscoring the importance of synthetic datasets in future AI projects. This phenomenon is encouraging ethical AI methods by employing artificial data to reduce prejudice and ensure fairer outcomes in AI uses.
"By Region, North America to have the largest market share in 2024, and Asia Pacific is slated to grow at the fastest rate during the forecast period."
In 2024, North America is expected to dominate the AI training dataset market with the largest market share. The reason for this dominance is the existence of big tech firms, significant investments in AI, and a strong network of data-centric advancements. Companies in North America are increasingly integrating artificial intelligence to enhance their operations, leading to a demand for high-quality training data. In the meantime, it is expected that the Asia Pacific region will show the highest rate of growth in the predicted period. The rapid expansion is due to additional investments in AI, higher internet usage, and a growing number of AI and machine learning startups. China and India are leading the way in embracing AI technologies, thanks to their abundant data and young population well-versed in technology.
Breakdown of primaries
In-depth interviews were conducted with Chief Executive Officers (CEOs), innovation and technology directors, system integrators, and executives from various key organizations operating in the AI training dataset market.
- By Company: Tier I - 18%, Tier II - 52%, and Tier III - 30%
- By Designation: C-Level Executives - 42%, D-Level Executives - 36%, and others - 22%
- By Region: North America - 42%, Europe - 26%, Asia Pacific - 21%, Middle East & Africa - 4%, and Latin America - 7%
The report includes the study of key players offering AI training dataset solutions. It profiles major vendors in the AI training dataset market. The major players in the AI training dataset market include Google (US), IBM (US), AWS (US), Microsoft (US), NVIDIA (US), Snorkel (US), Gretel (US), Shaip (US), Clickworker (US), Appen (Australia), Nexdata (US), Bitext (US), Aimleap (US), Deep Vision Data (US), Cogito Tech (US), Sama (US), Scale AI (US), Lionbridge Technologies (US), Alegion (US), TELUS International (Canada), iMerit (US), Labelbox (US), V7Labs (UK), Defined.ai (US), SuperAnnotate (US), LXT (Canada), Toloka AI (Netherlands), Innodata (US), Kili technology (France), HumanSignal (US), Superb AI (US), Hugging Face (US), CloudFactory (UK), FileMarket (Hong Kong), TagX (UAE), Roboflow (US), Supervise.ly (Estonia), Encord (UK), TransPerfect (US), Keylabs (Israel), and Data.world (US).
Research coverage
This research report categorizes the AI training dataset Market by Offering (Dataset Creation and Dataset Selling), by Dataset

Creation (Dataset Creation Software, and Dataset Creation Services), by Dataset Selling (Off-The-Shelf (OTS) Datasets, and Dataset Marketplaces), by Annotation Type (Pre-Labeled Datasets, Unlabeled Datasets, and Synthetic Datasets), by Data Modality (Text, Image, Audio & Speech, Video and Multimodal), By Type (Generative AI and Other AI), by End User (BFSI, Software & Technology Providers, Telecommunications, Automotive, Media & Entertainment, Government & Defense, Healthcare & Life Sciences, Manufacturing, Retail & Consumer Goods, And Other End Users) and by Region (North America, Europe, Asia Pacific, Middle East & Africa, and Latin America). The scope of the report covers detailed information regarding the major factors, such as drivers, restraints, challenges, and opportunities, influencing the growth of the AI training dataset market. A detailed analysis of the key industry players has been done to provide insights into their business overview, solutions, and services; key strategies; contracts, partnerships, agreements, new product & service launches, mergers and acquisitions, and recent developments associated with the AI training dataset market. Competitive analysis of upcoming startups in the AI training dataset market ecosystem is covered in this report.

Key Benefits of Buying the Report

The report would provide the market leaders/new entrants in this market with information on the closest approximations of the revenue numbers for the overall AI training dataset market and its subsegments. It would help stakeholders understand the competitive landscape and gain more insights better to position their business and plan suitable go-to-market strategies. It also helps stakeholders understand the pulse of the market and provides them with information on key market drivers, restraints, challenges, and opportunities.

The report provides insights on the following pointers:

?□Analysis of key drivers (increasing demand for diverse and continuously updated multimodal datasets for generative AI models, rising demand for multilingual datasets for conversational AI, demand for high-quality labeled data for autonomous vehicles, and Increased used of synthetic data for rare event simulation), restraints (legal risks of web-scraped data due to copyright infringement and limited access to high-quality medical datasets due to HIPAA compliance), opportunities (growing demand for specialized data annotation services in diverse fields, synthetic data generation and privacy-preserving techniques for augmented training data, and creation of customized AI Datasets and specialized formats (3D, AR/VR) for Enterprise Solutions), and challenges (data quality and relevance issues like inconsistency, bias, keeping datasets up to date, and diverse dataset formats and inconsistent annotation practices may hinder integration and reliability).

?□Product Development/Innovation: Detailed insights on upcoming technologies, research & development activities, and new product & service launches in the AI training dataset market.

?□Market Development: Comprehensive information about lucrative markets - the report analyses the AI training dataset market across varied regions.

?□Market Diversification: Exhaustive information about new products & services, untapped geographies, recent developments, and investments in the AI training dataset market.

?□Competitive Assessment: In-depth assessment of market shares, growth strategies and service offerings of leading players like Google (US), IBM (US), AWS (US), Microsoft (US), NVIDIA (US), Snorkel (US), Gretel (US), Shaip (US), Clickworker (US), Appen (Australia), Nexdata (US), Bitext (US), Aimleap (US), Deep Vision Data (US), Cogito Tech (US), Sama (US), Scale AI (US), Lionbridge Technologies (US), Alegion (US), TELUS International (Canada), iMerit (US), Labelbox (US), V7Labs (UK), Defined.ai (US), SuperAnnotate (US), LXT (Canada), Toloka AI (Netherlands), Innodata (US), Kili technology (France), HumanSignal (US), Superb AI (US), Hugging Face (US), CloudFactory (UK), FileMarket (Hong Kong), TagX (UAE), Roboflow (US), Supervise.ly (Estonia), Encord (UK), TransPerfect (US), Keylabs (Israel), and Data.world (US) among others in the AI training dataset market. The report also helps stakeholders understand the pulse of the AI training dataset market and provides them with information on key market drivers, restraints, challenges, and opportunities.

Table of Contents:

SCOTTS INTERNATIONAL

AI Training Dataset Market by Dataset Creation (Data Collection, Data Annotation, Synthetic Data Generation), Dataset Selling (Off-the-Shelf Datasets, Dataset Marketplaces), Data Modality (Text, Image, Video, Audio, Multimodal) - Global Forecast to 2029

Market Report | 2024-10-24 | 447 pages | MarketsandMarkets

To place an Order with Scotts International:

- - Print this form
- - Complete the relevant blank fields and sign
- - Send as a scanned email to support@scotts-international.com

ORDER FORM:

Select license	License	Price
	Single User	$4950.00
	Multi User	$6650.00
	Corporate License	$8150.00
	Enterprise Site License	$10000.00
	VAT	
	Total	

*Please circle the relevant license option. For any questions please contact support@scotts-international.com or 0048 603 394 346.

** VAT will be added at 23% for Polish based companies, individuals and EU based companies who are unable to provide a valid EU Vat Numbers.

Email*		Phone*	
First Name*		Last Name*	
Job title*			
Company Name*		EU Vat / Tax ID / NIP number*	
Address*		City*	

Zip Code*		Country*	
		Date	2025-05-20
		Signature	