LABEL YOUR DATA

THE BUYER'S GUIDE TO DATA LABELING VENDORS

Table of Contents

A word from the CEO

“At Label Your Data, we believe that informed choices lead to successful projects. In this guide, we’ve leveraged our expertise to provide you with a comprehensive overview of the top data labeling vendors, including our own company.

After investing roughly 120 hours of research, we’ve compiled everything you need to know about pricing, capabilities, and other criteria influencing your decision.

We hope you’ll find all the information you need to select the best data labeling partner.”

Karyna Naminas,
CEO of Label Your Data 🤍

LABEL YOUR DATA

Who is this guide for?

According to the report, internal teams and specialized service providers are equally popular for data labeling. This underscores the vital role of vendors, who often bring specialized expertise and efficiency, making them a better choice despite the equal popularity.

Common approaches for data annotation, %

This comprehensive guide is tailored for:

C-level working on AI:

Easily navigate the best options to streamline your data labeling processes, saving you time and effort.

AI and ML Engineers:

Find the most suitable vendors to enhance your ML ops efficiency and data quality.

AI Product Managers:

Make informed decisions for faster time-to-market strategies by choosing the right data labeling partners.

Academic researchers:

Let the experts handle massive datasets and time-consuming labeling, so you can focus on groundbreaking research.

You realize you need an ML solution

↓

Your ML accuracy is bad

↓

You seek public datasets

↓

You organize in-house annotation

↓

Your in-house annotation is nor efficient

↓

You search for automated process

↓

You start researching data labeling vendors

LABEL YOUR DATA

How to choose a dataset labeling vendor?

By now, you've likely tried every method for managing your ML workflows: labeling the data yourself, creating in-house or freelancer labeling teams, or partnering with an outsourcing vendor. Or perhaps you're looking to skip those steps and go straight to pre-built labeling workflows.

Regardless, establishing an outsourcing contract requires thorough research. What if the vendor isn't compatible with your tools, or their pricing is unreasonable? In this guide, we aim to answer your questions and streamline your vendor research. It's your chance to review well-known companies and discover new ones.

In the next chapters, we'll provide a brief overview of the top data labeling vendors and then analyze them according to these key criteria:

1. Services & Products
2. Client Reviews
3. Annotation Process
4. Pricing Options
5. Data Annotation Tools
6. Dataset Types
7. Security and Data Compliance
8. Quality Assurance
9. Integrations

For each criterion, we'll analyze how each vendor performs.

LABEL YOUR DATA

Overview of top data labeling companies

LABEL YOUR DATA	A service company offering a free pilot. There's no monthly commitment to data volume. Pricing calculator is on the website.
SUPERANNOTATE	A product company offering a data annotation platform. Provides a free trial and features a large marketplace of vetted annotation teams.
SCALEAI	A service company providing large-scale annotation solutions with flexible commitments. Offers transparent pricing options.
KILI TECHNOLOGY	A product company delivering a versatile data labeling platform. Features customizable workflows and powerful collaboration tools, with flexible pricing.
SAMA	A service company specializing in data annotation with scalable solutions, offering flexible pricing plans and focusing on social impact.
HUMANS IN THE LOOP	A service company providing expert annotation services for various industries. Offers flexible pricing plans and accurate, detailed annotations.
iMERIT	A service company offering end-to-end data annotation services with a global team. Provides scalable solutions and transparent, tailored pricing.
CLOUDFACTORY	A service company combining scalable data labeling with flexible pricing. Offers a free pilot to evaluate services before committing.

Feel free to screenshot this page and share it with your co-workers to make an informed decision.

LABEL YOUR DATA

Criterion 1: Services & Products

LABEL YOUR DATA

Our core services include:

Data labeling for computer vision

- Bounding boxes
- Optical Character Recognition
- Object and action detection
- Polygons
- Key points
- Semantic segmentation
- 3D cuboids
- LiDAR annotation

Data labeling for NLP models

- Text classification
- Sentiment analysis
- Named entity recognition (NER)
- Sentiment tagging
- Linguistic tagging
- Audio-to-text transcription

Additional services:

Quality Assurance (QA)

Ensuring the integrity of your datasets with our rigorous quality checks.

Model validation services

Identifying and addressing performance issues in your object recognition models.

Data collection services

For different types of data, depending on the goals of your ML algorithm.

Data entry

Capturing information from various sources and transcribing it into digital format.

Content moderation

Ensuring fairness and trust in maintaining your online platforms.

Data processing

Get all your data needs handled across any industry, from databases, unstructured data, to forms processing.

Industries we provide services to:

Academia / Agriculture / Aviation / Drones / E-commerce / FinTech / FinTech / Geospatial / Healthcare / Insurance / Manufacturing / Retail / Robotic

LABEL YOUR DATA

SUPERANNOTATE

They provide a platform to streamline the development of ML models. Its services include:

DATA ANNOTATION

FineTune: Creates training data for vario

WForce: Provides access to over 400 annotation teams spread across North and South America, Europe, Asia, and Africa. Subject-matter experts (SMEs) are provided for complex use cases.

DATA MANAGEMENT

Explore: Enables data management, version control, and debugging, facilitating the creation of accurate datasets efficiently. Users can visualize data trends and distributions to assess dataset health and annotator performance.

MLOPS & AUTOMATION

Orchestrate: Empowers users to build robust CI/CD pipelines for ML projects. Includes functionalities like built-in neural networks, Python SDK, webhooks, and advanced orchestration features.

LLMS & GENAI SERVICES

Services cover both SuperAnnotate software and expert workforce that are specifically designed to deliver quality training data for LLMs.

SuperAnnotate serves the following industries:

- Agriculture
- Robotics
- Healthcare
- Aerial imagery
- Insurance
- Sports
- Security and surveillance
- Autonomous driving

Their product solutions include a free desktop app called SuperAnnotate Desktop that addresses the slowness and lack of functionalities in existing annotation tools. It provides advanced features to speed up the labeling process.

LABEL YOUR DATA

SCALE AI

With a human-in-the-loop approach, the company implements the combination of machine learning and human input approach to provide their services:

DATA ANNOTATION

Scale AI provides annotation of images, videos, texts, maps, 3D images.

DATA CURATION

Involves testing, evaluating models, and comparing model tools to provide labels on only important objects and areas for model training.

REINFORCEMENT LEARNING FROM HUMAN FEEDBACK (RLHF)

The AI matter experts assess the model prompts and mark their outputs, evaluating their performance against the defined benchmark.

MODEL EVALUATION

Scale AI uses a red teaming approach to identify the model's risks and vulnerabilities. They use both LLM techniques and human insights.

GENERATIVE AI DATASETS CREATION

Their experts create high-quality datasets customized per your model and project. .

Scale AI's products include:

Scale Data Engine

Even though Scale AI positions it as a product, Data Engine is a process that improves the performance of ML models. It covers the end-to-end preparation of a model.

Scale GenAI Platform

This platform allows creating your own LLM based on your data and without need to send it to a third party. It has the necessary integrations and provides infrastructure to train, host, and evaluate a Gen AI model.

Scale Donovan

With the application of a LLM, Donovan helps with extraction and processing of tons of existing data. Donovan helps to segment and organize huge datasets for further ML processing.

KILI TECHNOLOGY

Their offerings include a data labeling platform, professional services with a global annotation workforce, and expert guidance from Machine Learning Engineers (MLE).

MANAGED LABELING SERVICES

Kili Simple

A global network of annotators who handle large-scale data labeling tasks.

ML Expert Guidance

Allows to hire annotators expert in over 100 domains and fluent in 30 languages to work within the platform.

Professional Services

Consulting services provided by Machine Learning Engineers (MLEs).

Kili has recently launched their new product, davinci, an AI tool powered by generative AI technology. It functions as a patent copilot, drastically accelerating patent drafting and office action response times compared to traditional methods.

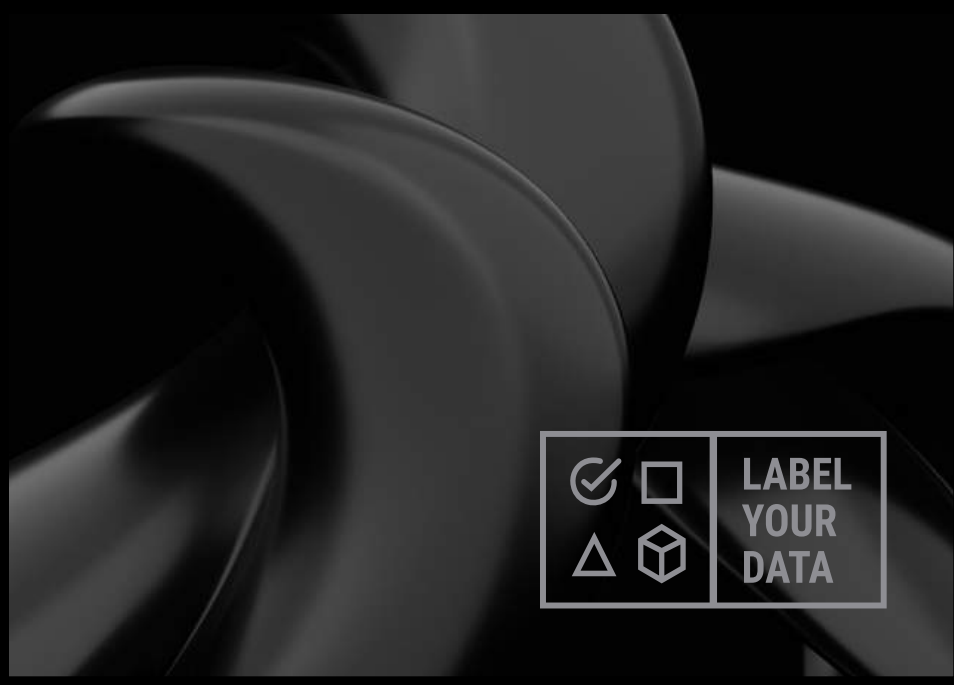
PLATFORM

A self-annotation solution that helps identify and fix inconsistencies within ML datasets. It offers AI-assisted tools that complement manual labeling.

- Labeling tools
- Quality management
- Seamless integration
- LLM Fine-Tuning
- LLM Evaluation & Testing

ASSETS

- Text annotation tool
- Image annotation tool
- Video annotation tool
- OCR annotation tool
- Geospatial annotation tool



SAMA

Sama leverages human-in-the-loop expertise within its labeling platform to deliver annotated data and help enterprises de-risk their ML model development. They specialize in various data annotation services for 2D and 3D images, videos, LiDAR, sensor fusion, and generative AI.

They also offer validation services for complex ML algorithms. Their services cater to industries like automotive, robotics, agriculture, retail, consumer tech, and manufacturing.

PRODUCTS

The main product is Sama Platform, an end-to-end data labeling and validation solution. This integrated ML-powered platform offers functionalities across the entire data annotation lifecycle:

- Data Pre-processing
- Task Management and Distribution
- Annotation Tools
- Quality Management
- Post-processing
- Task Delivery

SERVICES

Sama Curate

This AI-powered tool identifies assets needing labeling within your datasets. It works on both pre-filtered and completely unlabeled data.

Sama Annotate

A team of trained specialists providing data annotation for images, videos, and 3D point cloud data.

Sama Validate

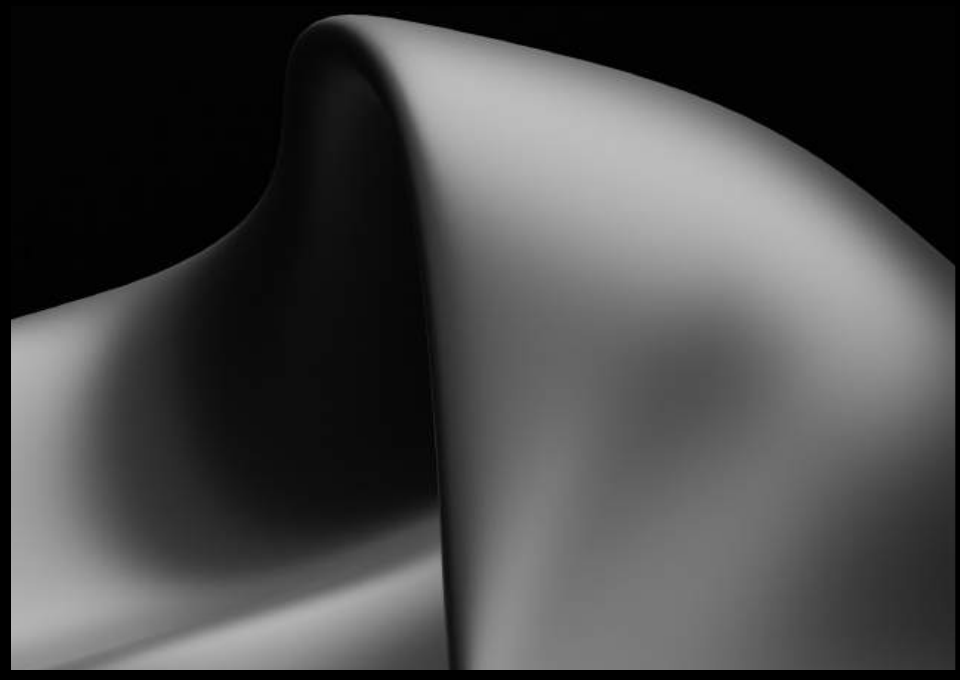
This managed service helps ensure enterprise AI models' effectiveness by reviewing predictions and making necessary corrections.

Sama GenAI

This solution addresses your generative AI needs: model evaluation, adversarial testing, training data preparation, and fine-tuning.

HUMANS IN THE LOOP

The company’s main expertise is computer vision, according to their website offerings. As such, Humans in the Loop doesn’t fit businesses that look for data labeling services for NLP models.



SOLUTIONS

Dataset annotation

They work with different types of data annotation, including bounding boxes, polygons, video annotation, keypoint, semantic, and 3D annotation.

Dataset collection

If you don’t have the dataset ready for annotation, the company collects the data for you. Their team can also help you expand the existing dataset to make it more detailed and diverse.

Human-in-the-loop for active learning

They can review your models to detect any anomalies or inconsistencies in its performance. They label the datasets, establishing new ground truth, which assists your model's further training.

Real-time edge case handling

Humans in the Loop serves as an intermediary between the model and the end user. The team accomplishes real-time AI model validation, reducing false responses.

Reinforcement learning with human feedback

This service focuses on ready models that need to be tested and improved before the final deployment. Their annotators review the model’s outputs, generating various questions in different languages to define your model’s weak points and signal them for further improvement.

LABEL YOUR DATA

iMERIT

PRODUCTS

Ango Hub

An AI-powered annotation platform for images, videos, and text, offering features like auto-labeling and tools for radiology, in-cabin monitoring, and defect detection.

SERVICES

Computer Vision Annotation

Image annotation

Bounding boxes, keypoint annotation, polygon annotation, image classification, semantic segmentation, and LiDAR.

Video annotation

The team works with bounding-box annotation, polygon, keypoint, and semantic segmentation annotation.

Image segmentation

Bounding boxes, grayscale, segmentation masks, Gaussian blur, among others.

NLP Annotation

Text annotation

Sentiment analysis, intent analysis, named entity recognition (NER), and entity classification.

Audio transcription

They convert audio data into text, and then label it for further ML processing.

3D sensor data

LiDAR annotation

Semantic segmentation, landmark annotation, 3D cuboids/box annotation, polygon and polyline annotation.

RLHF for LLMS and LVMS

iMerit provides domain expertise, expert feedback, and scalable experts-in-the-loop.

Product categorization

Puts images, videos, and text into defined categories for use cases like new product suggestions, query-understanding algorithms, and personalized recommendations.

Content moderation

Helps monitor, assess, and filter various user-generated content (UGC), detecting offensive content.

LABEL YOUR DATA

CLOUDFACTORY

By bringing together people and technology, delivering human-in-the-loop AI solutions, including:

SERVICES

Data labeling

Accelerated Annotation

AI-powered labeling for 2D images and videos, achieving up to 30 times faster labeling.

Workforce Plus

A complete package for labeling video, LiDAR data, and more, along with the tools you need.

Vision AI Managed Workforce

A dedicated workforce trained for computer vision tasks (Vision AI).

NLP

Text and audio data labeling tasks.

Data Processing

Data processing and other back-office tasks.

Human-in-the-Loop Automation

Managed Workforce

A skilled workforce that complements AI automation

PRODUCTS

Hasty (acquired in 2022)

A data-centric ML platform specifically designed for creating and deploying computer vision applications. It offers features like AI-powered image annotation, quality control, and no-code model building.

The platform can handle:

- Classification
- Tagging
- Object Detection
- Instance Segmentation
- Semantic Segmentation
- Panoptic Segmentation
- Attribute Prediction

AI assistants

AI tools within Hasty to automate some or even all of your annotation tasks.

CloudFactory's data labeling services primarily target computer vision tasks across various industries but offer limited support for non-English NLP tasks.

LABEL YOUR DATA

APPEN

Appen provides a comprehensive suite of annotation services. The company also offers an end-to-end platform to support the entire AI development lifecycle:

SERVICES

Data Collection

Appen sources various data types, including text, audio, video, and geospatial data.

Data Annotation

Human-in-the-loop annotation for NLP, speech processing, and computer vision.

Search Relevance

Enhancing search engine algorithms through model evaluation, content moderation, and related search refinement.

Reinforcement Learning

RLHF services specifically for developing LLMs.

Document Intelligence

Improves document processing through data curation and annotation.

Data for LLMs

Custom data solutions for LLM builders.

Pre-Labeled Datasets

A library of 270+ pre-labeled audio, image, video, and text datasets in over 80 languages.

Location-Based Services

Geospatial data annotation for accurate mapping and location intelligence.

PLATFORM

A global crowdsourcing network with over 1 million contributors across 170+ countries. It offers features like project management, performance monitoring, data quality control, and security management. It also provides access to subject-matter experts (SMEs).

The platform handles various tasks:

- Data Collection
- Data Annotation
- Transcription
- Translation
- Speech Modeling
- Model Evaluation

LABEL YOUR DATA

Criterion 2: Client reviews

Selecting a data labeling vendor hinges not just on features, but also on real-world performance. We'll present you the summary of client reviews on top platforms like G2 and Clutch to help you assess vendor quality through the lens of their past customers.

Clutch Reviews

Label Your Data	★★★★★ **5.0**

As of 2024 we are the only company that has Clutch-vetted B2B reviews. Companies missing from this table are not on Clutch.

G2 Reviews

Label Your Data	4.9
SuperAnnotate	4.9
Scale Document AI	4.9
iMerit Ango Hub Platform	4.8
Kili Technology	4.7
Sama	4.6
CloudFactory	4.5
Scale Rapid	4.4
Appen	4.2
Humans in the Loop	0

Criterion 3: Annotation process

The most crucial factor in evaluating data labeling companies for your ML project is understanding how they label your data step-by-step. This directly impacts the quality and consistency of your training data, which ultimately determines your model's success.

LABEL YOUR DATA

We deliver a streamlined and secure annotation process for your ML projects. Here's how our collaboration will look like:

1 Run Free Pilot

We deliver a streamlined and secure annotation process for your ML projects. Here's how our collaboration will look like:

2 Get Cost Calculations

After the pilot, you receive a comprehensive cost estimate for annotating the remaining objects. This ensures transparency and allows you to make an informed decision based on our actual performance.

3 Sign Confidentiality Agreements

Your data security is our priority. We are committed to safeguarding your information. Signing a non-disclosure agreement (NDA) is a mandatory step before commencing any project.

4 Experience Gradual Dataset Delivery

Receive the first batches of annotated datasets promptly, enabling you to initiate the machine learning model training process.

5 Benefit from Iterative Refinement

With each training iteration, you'll benefit from an increasingly accurate model as our annotations continue to refine its performance.

This structured approach ensures a collaborative and secure annotation experience.

LABEL YOUR DATA

SUPERANNOTATE

SuperAnnotate offers various project types within their platform. Here's how the annotation process looks like:

1. Choosing the project type

Leverage the platform for various project types: use AI tools for image segmentation and text extraction, track objects in videos, classify text snippets, annotate high-resolution images, and create projects with custom interfaces for LLMs and generative AI.

2. Setting up the project

Leverage the platform for various
Create a new project by choosing the project type that aligns with your data and give it a clear and descriptive name. After that, proceed to data import.

3. Annotating data

SuperAnnotate provides tools specific to your project type. Use these tools to label and define the objects within your data.

4. Utilizing classes and attributes

SuperAnnotate provides tools specific to your project type. Use these tools to label and define the objects within your data.

In terms of annotation services, the vendor follows a structured process:

Project management

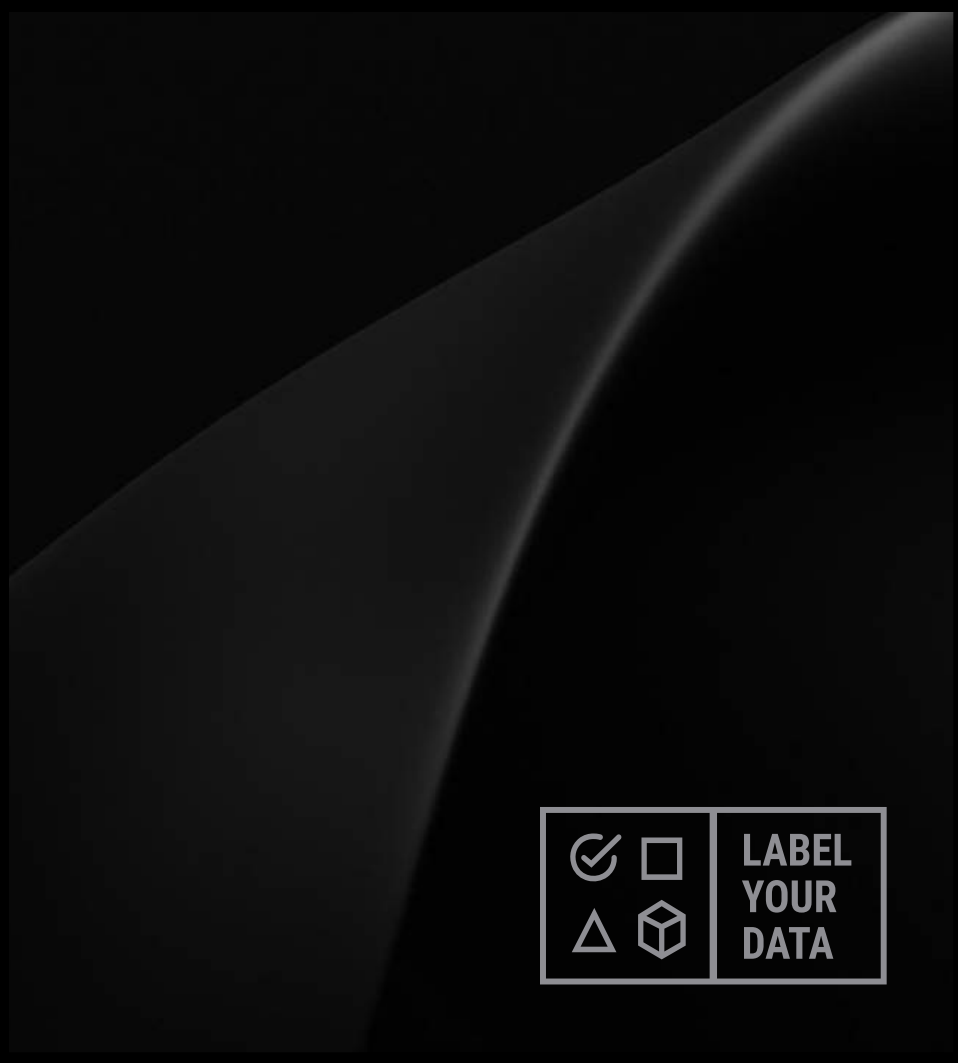
A dedicated Project Manager (PV) will oversee your team of annotators and quality assurance (QA) specialists.

DataOps

Your PM takes care of importing your dataset: setting up the relevant classes, adding clear instructions for the annotators, and exporting the completed project.

MLOps

The labeled data you create is seamlessly integrated into your ML pipeline. This allows you to use the labeled data to train your ML models effectively.



SCALE AI

Since Scale AI uses its own platform for data annotation, the process is quite smooth:

Uploading data

First, you need to upload the datasets. By using the platform, you choose one of the suitable formats (attachment from computer, link, sharing through cloud storage, etc.).

Setting project goals

You provide detailed instructions for labeling and set the benchmarks that correspond to the desirable annotation. You also specify the desirable number of reviews needed per task.

Annotating data

Data annotation is done in three pipelines: the standard pipeline with one review attempt, the consensus pipeline with three attempts, and the beta collection pipeline, which provides all annotator responses without consolidation.

Exporting data

Once the annotation is complete, you receive the final output and download the datasets through Scale API.

KILI TECHNOLOGY

The company provides a variety of services and approaches to data labeling, which depends on the specific service you choose:

Platform-Based Annotation

- Build customized interfaces
- Streamlined labeling & analysis
- In-depth review & refinement
- Seamless cloud integration & access control
- Enhanced workflow & automation

Kili Simple Annotation

- Upload data sample
- Quote and project start
- Monitor progress
- Receive a fully labeled dataset

Kili Simple Annotation

- Communicate project requirements
- Provide feedback and validate quality
- Monitor progress and refine annotations
- Continuous improvement with active learning

LABEL YOUR DATA

SAMA

Sama offers a secure cloud-based platform to manage the entire image annotation lifecycle. A team of trained specialists is assigned to your project, including annotators, project managers, engineers, and quality analysts.

1. Consultation

A dedicated account team collaborates with you to design a robust data quality strategy and annotation workflow.

2. ML-powered Platform

Annotators leverage Sama's ML-powered platform to streamline the annotation process. This may involve features like pre-suggested labels or automated quality checks.

3. ML-powered Platform

A multi-layered quality control process includes the review performed by humans and potentially automated features like Auto QA to identify and address errors or inconsistencies.

4. Delivery and Support

Sama offers detailed analytics and customized reporting on your training data, with a CLI and APIs to automate data transfers, task prioritization, and real-time results retrieval.

HUMANS IN THE LOOP

When collaborating with Humans in the Loop, you go through a number of steps:

Introduction call

During the introduction call, you describe your project needs and requirements in terms of geographic location, demographics, and types of annotation.

Free trial

A free trial consists of 2 hours of free annotation on the sample and according to the instructions you will send to them. The free trial is usually ready 72 hours after you send your samples.

Fine-tuning

Humans in the Loop reviews your task and specify the adjustments, if required. You agree on the annotation type and complete the interface set-up.

Contract conclusion

As soon as you sign the contract, a certain amount of annotators will be assigned to your project. Project coordinators will take care of the milestones. The team can provide a custom report to track the progress.

Project delivery

The delivery typically takes place through the API interface you set up previously. The company also sets up the final call to ensure the task is complete.

LABEL YOUR DATA

iMERIT

Here are their main steps for data annotation at iMerit:

Consultation with an expert

You register with iMerit's platform and prepare your task.

Trial and annotators' training

Depending on the task, annotators proceed with a pilot or proof of concept. The group of annotators is chosen and trained, especially if the task requires some specific industry knowledge.

Workflow customization

Data annotation happens for the piece of the project, which is defined in the pilot stage.

Feedback cycle

The client provides feedback and proceeds to the offer.

Evaluation

At the end of every project, the evaluation is done before submission of the final project.

You may be involved in the annotation process at any stage you decide.

CLOUDFACTORY

Instead of having a traditional data labeling process, CloudFactory takes an integrated approach:

Free analysis

They don't offer a traditional free trial, but they do provide a complimentary "analysis" of your project. It functions as a mini pilot project, taking around 10 hours to complete.

Team onboarding

They take two weeks to onboard a dedicated team for your project to meet your service level agreement (SLA). If necessary, they'll also recruit additional workers during this time.

Data annotation

Once the team is ready, they'll begin labeling your data according to your specifications.

Quality assurance

Every step involves rigorous quality checks to ensure accuracy.

Process iteration

They monitor the process and make adjustments as needed. This could involve refining data features, adapting task workflows, or enhancing QA procedures.

Project management

They handle project planning, process implementation, and ongoing measurement to ensure your project meets the desired outcomes.

CloudFactory provides a dedicated Client Success manager and Delivery Team Lead for each project, as well as a dedicated Channel Manager for ongoing support.

APPEN

Onboarding and training

A dedicated Customer Success Manager gets your team familiar with the platform, ensuring they can quickly create and launch annotation jobs.

Setting up annotation jobs

This involves defining the type of data and the specific annotations required. Jobs can be set up directly through Appen's user interface or their API.

Data annotation by global contributors

Appen sends your jobs to a global network of contributors qualified for the specific task.

Monitoring and adjustment

You can monitor the progress of your jobs and review incoming data to see if adjustments are needed.

Data download and reporting

Once the annotation is complete, you can download the labeled data. Appen provides dashboards and reports to help you optimize your jobs for factors like cost, data quality, and efficiency.

LABEL YOUR DATA

Criterion 4: Pricing options

COMPANY	PRICING
LABEL YOUR DATA	• Keypoints = $0.015 • Rectangles = $0.02 • Polygons = $0.045 • Cuboids = $0.09 • Entities to label = $0.02 • Characters in a dataset = $0.0005 • Annotation hours = $6.0 • Other →
SUPERANNOTATE	• Cost-per-unit pricing with quality benchmarks • Options include Tool Purchase and All-in-One Service • A 14-day trial of Pro Plan to explore the tool's functionalities
SCALE AI	• Custom pricing for enterprises • Pay-as-you-go model for individuals • 1000 labeling units are offered for free • Specific costs per task for Scale Rapid
KILI TECHNOLOGY	• Tiered pricing: Free Plan, Grow Plan , and Enterprise Plan • Free Plan is limited to 5,000 annotations and 5 collaborators.
SAMA	• Per-feature pricing • Contact directly for a quote for specific project requirements
HUMANS IN THE LOOP	• Custom pricing based on individual client needs • Calculated per task or per hour • Offers express annotation services.
İMERIT	• Monthly subscription for platform usage • Pricing varies by dataset volume • Discounts for large volumes and additional charges for custom exports
CLOUDFACTORY	• Per object for computer vision tasks • Hourly for NLP tasks • Yearly agreements and discounts for high-volume projects
APPEN	• Flexible, unit-based, and hourly pricing • Free trial • Formula based on work complexity and volume

Data labeling companies offer various pricing options to fit project needs. There are a few critical factors for you to consider that influence the labeling price:

- **Project duration:** long-term or a one-time project
- **Quality, cost, and turn-around time:** rank these in order of importance
- **Pricing model:** per-label, hourly rates, or tiered subscription plans
- **Internal costs:** consider your own budget and resource limitations

Criterion 5: Data annotation tools

The next step in choosing a data labeling vendor is evaluating their core tools, proprietary solutions, and LLM integration for automation.

Product companies usually limit labeling to their own tools, reducing flexibility, while service companies like Label Your Data offer adaptable solutions that integrate with various tools and workflows.

LABEL YOUR DATA

VENDOR	ANNOTATION TOOLS	LLM AUTOMATION
LABEL YOUR DATA	CVAT, Labelbox, Label Studio, SuperAnnotate, Datature, Supervisely, QGIS, V7, doccano, Philosys (LiDAR), BasicAI (LiDAR), ubiAI Adapts to client-preferred tools or custom-built solutions	**Custom**
SUPERANNOTATE	LLM Annotation Tool, Image Annotation Tool, Video Annotation Tool, Text Annotation Tool, Audio Annotation Tool, Classification Tool Automated platform features, SAM integration Pre-labeling feature under development	✓
SCALE AI	Scale Rapid, Scale Studio	✓
KILI TECHNOLOGY	Text Annotation Tool, Image Annotation Tool, Video Annotation Tool, OCR Annotation Tool, Geospatial Annotation Tool Uses built-in AI models, like ChatGPT and SAM, to automatically pre-label data	✓
SAMA	ML-Assisted Annotation (MAA), Crosshair Tool, Visual Feedback on Label Selection, Logically Grouped Tasks, Keyboard Shortcuts, SamaIQ, SamaHub	✓
HUMANS IN THE LOOP	Partnerships with: Alegion, Diffgram, Human Lambdas, Hasty, Kili Technology, Lightly, V7, Manthano, Superannotate, Segments.ai To use your specific tool, you will pay an extra fee.	✗
IMERIT	Ango Hub Can use client's tools, but prefer their own.	✓
CLOUDFACTORY	CloudFactory platform Partnerships with: Dataloop, Datasaur.ai, Labelbox, custom tools	✗
APPEN	Appen's Data Annotation Platform (combination of human annotators and machine learning), client-provided tools integration	✓

Criterion 6: Dataset types

The second crucial aspect of finding the right data labeling partner is analyzing the types of data they can handle. Ideally, the vendor should offer expertise in both Computer Vision and NLP. This flexibility ensures they can adapt to your project's evolving needs and diverse data formats.

LABEL YOUR DATA

We handle a wide range of dataset types and formats to ensure your ML project runs smoothly:

Images & Videos: Common image formats like JPEG, PNG, TIFF, and BMP, as well as video formats like MP4, AVI, and MOV for image classification, object detection, and video analysis.

Text Data: TXT, DOCX, PDF, and spreadsheet files (.CSV, .XLS) for sentiment analysis, text classification, and NER. For text projects, we can also handle code files where annotations are made directly on the text within the code.

Audio Files: We can process audio data in formats like MP3, WAV, and FLAC for audio transcription and sentiment analysis.

LiDAR Data: We can handle LiDAR data formats for 3D object detection and scene reconstruction tasks. Primarily, we work with LiDAR data in the PCD format. However, we can also accommodate client-provided LiDAR data in JSON format, commonly used for calibration information and configuration files.

Custom Data Formats: For unique project requirements, we can work with custom formats and develop solutions to integrate your data into our labeling workflow.

SUPERANNOTATE

The SuperAnnotate teams work with various data types, including images, audio, videos, LiDAR, text, and even custom data formats. Their platform offers support for the following data types:

Images. SuperAnnotate supports common image formats including JPG, JPEG, PNG, WEBP, TIFF, BMP, and specific TIF variations.

Videos. For video annotation, SuperAnnotate works with MP4, AVI, MOV, FLV, MPEG, and WEBM formats within image projects. Video or audio-specific projects can utilize OGG, WEBM, and MP4 formats (compatible with HTML5).

Text. Plain text files in TXT format with UTF encoding are also supported for annotation tasks.

Tiled Imagery. SuperAnnotate allows for the management of tiled imagery, which breaks down large images into manageable sections for annotation.

Point Cloud Data. For 3D applications, SuperAnnotate can handle point cloud data, a format representing 3D objects using points in space.

SCALE AI

The company works with varied dataset types:

Text. Various types of documents and transcriptions. They work with NLP models and different content. You can annotate datasets for content classification, text generation, transcription, content collection, and NER.

Image. They can process electro-optical imagery, infrared, and transcription images. The annotation is possible with bounding boxes, polygons, key points, ellipses, cuboids, and lines.

Audio. Diverse audio datasets for annotation, both from active and passive sonars. They work with entities for the same cases as with the text.

Video. Their specialists prepare the models for the natural language processing, annotating full motion videos.

3D sensor fusion. Providing LiDAR annotation and map labeling, offering services to autonomous driving industry, robotics, and virtual reality.

KILI TECHNOLOGY

Kili's platform streamlines the annotation process for various unstructured data formats. You can annotate images, videos, text documents, PDFs, satellite imagery, and even conversational data.

Moreover, the vendor makes it easy to transfer your data. You can import your data for annotation and then export the annotated data to use for further analysis or model training. The supported formats for import are CSV, JSON, and image files. For export, you can choose between CSV, JSON, and TensorFlow Record.

However, we've discovered a few platform limitations:

- **Video annotation:** Currently, the platform supports only bounding boxes.
- **OCR integration:** Text annotation in images requires users to upload their own OCR data, which adds an extra step to the workflow.

Supported formats:

Excel and Word documents need to be transformed before import.

SAMA

Sama company supports several dataset types for computer vision annotation projects only:

- **Image datasets.** The datasets that contain still images that Sama can annotate with various methods, including bounding boxes, polygons, keypoints, segmentations, and lines and arrows.
- **3D sensor fusion datasets.** This involves combining data from multiple 3D sensors like LiDAR and cameras.
- **Video datasets.** Sama can also handle video annotation projects.

Overall, Sama specializes in annotating various data formats, from 2D images to complex 3D sensor fusion projects. However, you should keep in mind that the vendor works only with data labeling for computer vision tasks.

HUMANS IN THE LOOP

Human in the Loop AI model preprocessing considers various formats. This relates both to data collection and data annotation. The company works with the most popular annotation formats, including JSON, Yolo, COCO, Pascal Voc XML. At the same time, the specialists remain open and flexible for other cooperations. They can consider your format for annotation or data export.

İMERIT

The company works with a wide variety of dataset types. They are able to prepare datasets for computer vision, sentiment analysis, natural language processing, categorization, and LiDAR annotation.

The majority of their annotation types include polygons, bounding boxes, keypoints, polylines, classification, semantic segmentation, text extraction, and others. Besides, they can perform audits and quality assurance (QA) of generative AI systems.

APPEN

Appen caters to a wide range of ML needs by offering various dataset types. Text data, encompassing emails, documents, and chat conversations in over 80 languages, can be processed and annotated. They also handle audio data, like phone calls and voice commands, providing transcriptions and labels for aspects like speaker identification.

Appen's expertise extends to visual data as well, including image annotation for objects and scenes, along with facial recognition. Even video data, from security cameras to user-generated content, can be analyzed for object tracking and activity recognition.

Supported Data Formats:

CSV, TSV, XLSX, and ODS.

Encoding:

All files must be saved using UTF-8 encoding.

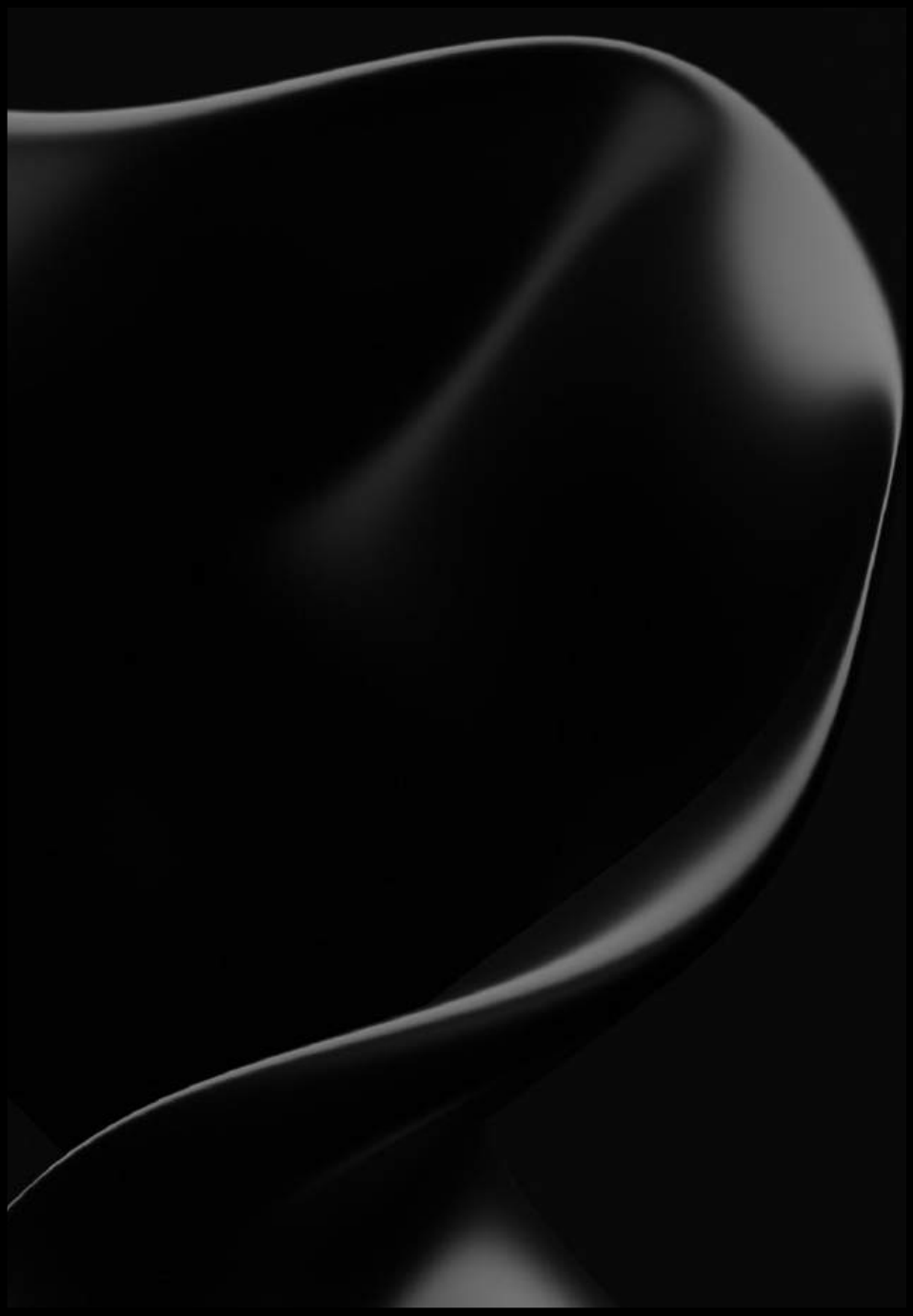
Formatting:

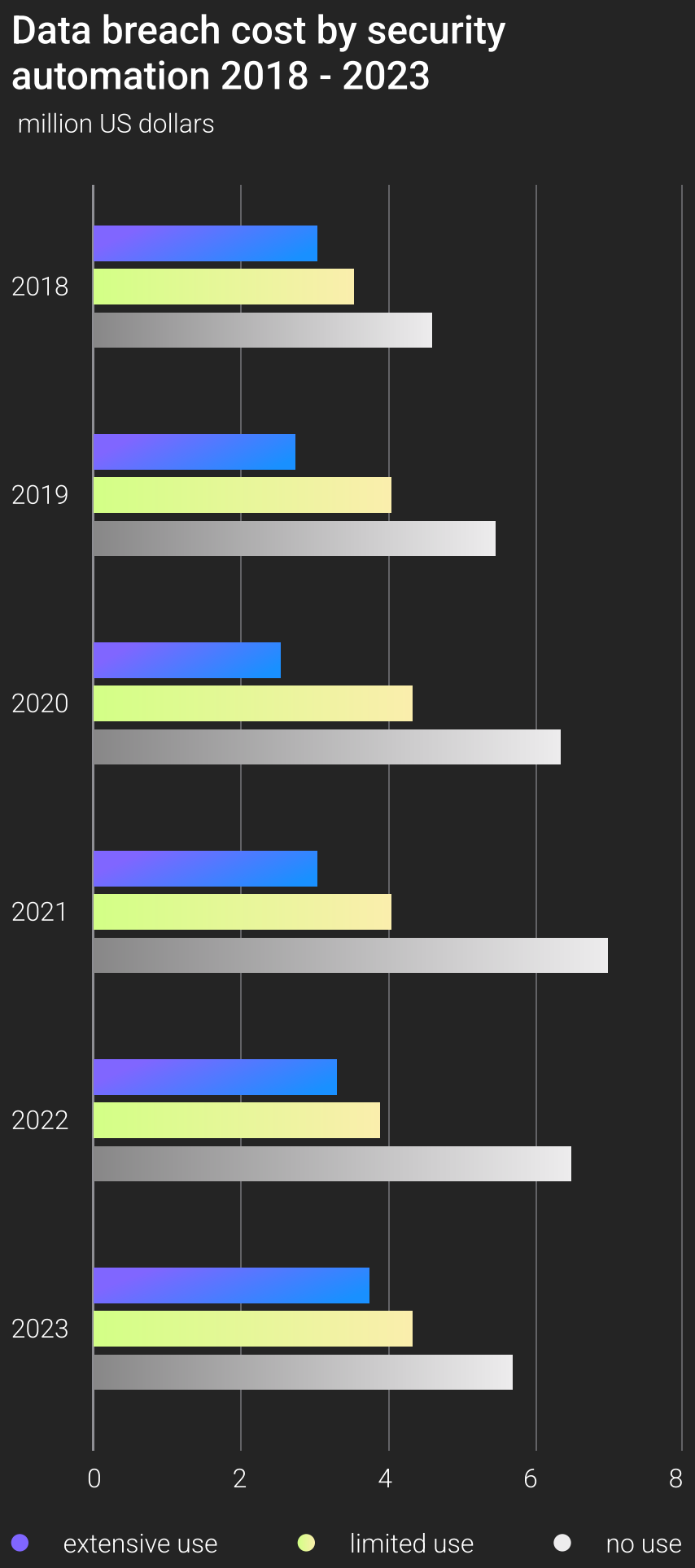
Each column in your data file must have a clear and descriptive header.

Criterion 7: Security and data compliance

The most crucial factor in evaluating data labeling companies for your ML project is understanding how they label your data step-by-step. This directly impacts the quality and consistency of your training data, which ultimately determines your model's success.

Choose a company with robust encryption, access controls, and relevant compliance certifications to protect your sensitive information.





LABEL YOUR DATA

COMPANY	SECURITY CERTIFICATES	DATA COMPLIANCE
LABEL YOUR DATA	• PCI DSS Level 1 • ISO/IEC 27001:2013 • GDPR • CCPA • HIPAA	• Data leaks prevention • Team screening • Secure workspace • Robust infrastructure • Continuous training • Secure software practices • NDAs • Encryption
SUPERANNOTATE	• SOC 2 Type 2 • CCPA • GDPR • HIPAA • SSO • 2FA	• AWS S3 secure storage • Private VPC • Restricted access • Prompt deletion • Granular permissions • Robust access controls
SCALE AI	• SOC 2 Type II • HIPAA • DoD IL4 • ISO 27001 • FedRAMP (in progress)	• Robust access controls
KILI TECHNOLOGY	• ISO 27001:2013 • SOC 2 Type II • HIPAA	• Access controls • Encryption • Anti-malware, secure development practices • Incident management • Risk management
SAMA	• ISO 9001 • ISO 27001 • GDPR • TISAX • 2FA	• Automated security scanning • Encryption

HUMANS IN THE LOOP	• GDPR • SOC 2 • SSO • 2FA	• Security awareness training • Cloud security • Access security • Third-party audits • Annual risk assessment
İMERIT	• SOC 2 • ISO 27001 • ISO 9001:2015 • GDPR • TISAX	• Security Manager for each client
CLOUDFACTORY	• ISO 9001:2015 • ISO 27001:2013 • SOC 2 • HIPAA • GDPR	• Secure network environment • NDAs
APPEN	• GDPR • HIPAA • SOC 2 Type II • ISO 27001	• Temporary link for view-only access

Criterion 8: Quality assurance

The next factor to keep in mind is how strong is the QA process in a data labeling company. It ensures the accuracy and consistency of your data, leading to a well-trained machine learning model.

VENDOR	QA PROCESS
LABEL YOUR DATA	**Achieves 98% accuracy with:** ● Project requirement gathering ● Annotator training ● Pilot projects ● Cross-reference QA ● Random sampling ● Milestone-based quality control
SUPERANNOTATE	**Multi-stage review process:** ● Annotators complete initial labeling ● QAs review and approve annotations ● Project Admins have final oversight **Additional methods:** ● Vetted workforce ● Multi-level QA system ● Dedicated project management ● Separate QA teams
SCALE AI	**Review cycle with two annotation layers:** ● Annotators label data ● Second layer monitors and corrects ● Consensus pipeline for final version ● Quality screens with accuracy tests ● Evaluation tasks for benchmarking
KILI TECHNOLOGY	**Ensures quality through:** ● Annotation consistency checks ● Review and feedback loops ● Quality control metrics ● Continuous feedback ● Targeted reviews ● Programmatic QA

SAMA	**Three-way QA approach:** ● Internal QA with manual and automated techniques ● Client-driven QA with data scientist reviews ● Automated QA (Auto QA)
HUMANS IN THE LOOP	**Two-layer QA process:** ● Dedicated supervisors ● QA teams check before submission ● Annotators document edge cases
iMERIT	**Uses:** ● Reports and dashboards ● Gold standards during mini-sets ● Annotator consensus ● Scientific methods for label consistency ● Subsampling with random sample reviews ● AI-based frameworks
CLOUDFACTORY	**Combines automated checks and human review:** ● Built-in QA ● Model feedback **Multi-layered quality control with:** ● Gold standard ● Sample reviews ● Consensus ● Intersection over Union (IoU)
APPEN	**Customizable QA plans with:** ● Human expertise and AI options ● Seamless workflow integration ● Rigorous quality control with multiple checks ● AI-powered crowd management ● Detailed quality reports for tracking progress

Criterion 9: Integrations

COMPANY	INTEGRATIONS
LABEL YOUR DATA	AWS S3, API, custom storage solutions, public access URLs, custom integrations on demand.
SUPERANNOTATE	AWS S3, GCP Buckets, Azure Containers, Databricks, Python SDK, custom storage solutions.
SCALE AI	Public access URLs, AWS S3, Google Cloud Storage, Azure Blob Storage, Scale file upload API, IT whitelisting, Scale API, Sail SDK, Python SDK.
KILI TECHNOLOGY	Amazon S3, Google Cloud Storage, Microsoft Azure Blob Storage, API, Python SDK, webhooks, version control.
SAMA	REST API, Python SDK, CLI, Azure Blob Storage, Google Cloud Platform, AWS S3, Sama S3 Storage.
HUMANS IN THE LOOP	APIs, custom integrations on demand.
İMERIT	AWS S3, Google Cloud Storage, Azure Blob Storage, APIs, Webhook, Keras, PyTorch, TensorFlow, plugins, custom requests.
CLOUDFACTORY	AWS S3, Google Cloud Storage, Azure Blob Storage, REST API, machine learning frameworks (TensorFlow, PyTorch).
APPEN	API, Live LLM APIs, RESTful API with JSON data format

ⓘ This guide is based on our own research coupled with publicly available information on the vendors' websites: superannotate.com / scale.com / kili-technology.com / sama.com / humansintheloop.org / imerit.net / cloudfactory.com / appen.com.

LABEL YOUR DATA

Why Choose Label Your Data

Selecting a reliable data labeling partner for your ML project can be tough, but we hope that with this guide, you can make an informed decision.

We've put a lot of effort into matching you with the best vendor. But if you find our advantages the most appealing to your case, let's run a pilot project together.

Flexible pricing

Pay per labeled object or per annotation hour.

No commitment

Check our performance based on a free trial.

Data-compliant

Work with a data-certified vendor: PCI DSS Level 1, ISO:2700, GDPR, CCPA.

Tool-agnostic

Working with every annotation tool, even your custom tools.

Run free pilot!