MULTILINGUAL AI & DATA SERVICES
COLLECTION, ANNOTATION, VALIDATION

venga

Venga delivers custom solutions for AI training data sets. We evaluate each project for fit-for-purpose rules and our teams' guidelines. Our solutions are fully scalable, accurate, and adapted to your needs.

DATA COLLECTION

Custom text, voice, image, and video data in over 75 languages for your AI testing and training sets.

Text Data Collection

Clean, human-generated text data for machine learning systems.

Quality AI needs good, clean data sets. We collect data based on your system's unique requirements. With qualified resources around the globe, we rapidly scale according to your project and cover more than 75 languages. Our human-in-the-loop process provides better data than scraping or crowdsourcing alone. All of this means better performance for your machine learning systems.

Image & Video Data Collection

We capture video and image data for your object and sentiment recognition systems.

AI object recognition requires large image data sets that conform to your preferred labeling and annotation methods. Venga provides custom image capture and annotation services that include bounding boxes, sentiment analysis, handwriting recognition, and transcription.

Audio Data Collection

We collect multilingual audio and voice data for Automatic Speech Recognition (ASR) software.

ASR systems need large quantities of high-quality language data from numerous contexts and environments. Venga provides custom audio data sets that match your requirements for speaker profile, subject matter, and background sounds.

Curation

Sometimes scraping doesn't deliver the training data you need.

Do you need to test with data types that are too complex to scrape? We also curate fully customized data. Need bilingual sentences with inline HTML tags or emojis? How about one-minute videos with transcription and translation? Our solutions architects are ready to discuss your needs and provide proof-of-concept solutions.

Handwritten Data Collection

Train optical character recognition (OCR) with high-quality handwritten data collection.

Handwritten text recognition requires large data sets of text images paired with transcriptions and annotations. Venga provides clean text data in over 75 languages. Our human-in-the-loop process ensures that your training sets are clean and accurate, so your OCR systems learn better and faster.

DATA ANNOTATION

Data annotation is the human-powered labeling of text, images, video, and audio to allow machine learning systems to recognize objects. We preprocess your data with metadata tags to make clean, usable data sets for your AI training programs.

Text Data Labeling

Data labeling enriches your text with meta tags that provide the context, structure, and object recognition needed for AI training sets.

Text data labeling is a type of annotation in which meta tags identify parts of a text and add rich information for machine learning systems. These include linguistic annotation, entity annotation, and sentiment analysis. Venga discovers your custom labeling needs and provides human-powered annotations in over 75 languages.

Image & Video Annotation & Transcription

Image labeling for pattern recognition and computer vision systems.

When public data sets are not enough, Venga provides customized image annotation. We offer annotation services including bounding boxes, semantic segmentation, image classification, and text transcription. With a global team, Venga provides clean, multilingual data for your machine learning systems.

Audio Annotation & Transcription

Transform your audio data into a simpler format that can be read and parsed by AI through transcription and annotation.

Text transcriptions and descriptive meta tags for your audio make it more usable for machine learning. Transcriptions allow AI systems and search engines to crawl your audio and understand it. Annotation provides richer information for your machine learning models.

Audio Classification

Classify audio data for improved natural language processing (NLP) in speech recognition, chatbots, text to speech, and voice search.

Audio classification is the process of analyzing audio data and categorizing it for use in machine learning. With in-country resources, Venga meets your multilingual audio classification needs. Our human-based approach generates clean data sets to improve your natural language processing systems.

Sentiment & Intent Analysis

Categorize and annotate sentiment and intent in your text, voice, image, and video data with quality human analysis.

Intention and emotion can be a challenge for AI to understand. They often require large training sets of human-annotated sentiment data for reliable results. Such data sets include text analysis, social listening, emotion analysis, opinion mining, and language variations. Venga is a global company with in-country data analysis resourcing to cover your multilingual sentiment analysis needs.

DATA VALIDATION

The value of data sets decreases exponentially when the quality is poor. Validate your data to improve accuracy in your machine learning models.

Ad Evaluation

Optimize ads by gathering quality opinion data from your target demographics.

Before launching an expensive PPC campaign, we run your ad by your target audience to define sentiment, intent, and effectiveness. With in-country resources in over 75 languages, we optimize your ad for any global target market.

Data Redaction

Comply with privacy requirements and make the most of existing data sets by redacting sensitive or personal data.

Common types of data redaction include blurring or distorting of images, text, or audio. Venga has secure systems to redact your data and make it usable on a larger scale. With in-country resources all over the world, we can help you comply with local privacy laws when working with large AI training sets.

Content Moderation

Peace of mind with brand safety and security.

As the importance of online customer engagement increases, so does the risk of an inappropriate post or image damaging your brand's reputation. Venga provides content moderation services to identify potentially risky content and flag it. We can use a submission approval process, or we can scan your channels after content has been published. We have qualified analysts fluent in over 75 languages to keep your content culturally appropriate for any target market.

Search Evaluation

Ensure your search algorithms respond to queries correctly and that your website pages display correctly when searched.

You have a lot of content. Our multilingual team will audit your search results pages and identify areas for improvements including caption evaluation, ad review, and search query categorization. We customize each project to fit your unique needs.

MULTILINGUAL DATA

Translating data for machine learning requires more than just native speakers. Our linguists are trained in data collection translation projects as well. They adapt their translation style to the task and follow specified machine learning rules for quality data sets.

Data Translation

We make your data more usable with data translation for Natural Language Processing (NLP) training models.

Clean data sets are vital when training AI to respond to written and verbal language. Using supervised data collection, Venga delivers multilingual data sets of text, voice, and image data suitable for test and training sets. Venga has an extensive pool of qualified resources. We can scale to 1000+ linguists for quality data set translation in 2 to 4 weeks.

Machine Translation

Machine Translation (MT) offers fast turnaround times, large volumes, and low costs. Venga creates custom MT solutions to meet our clients' needs.

Venga provides a customized MT engine that gets smarter as it learns customer preferences. We work with you to determine solutions for your content from baseline engine selection to workflow setup. We have a detailed quality assurance process for optimizing time and cost savings at the right level of quality.

Machine Translation Quality Evaluation

Trained linguists review your machine-translated context to ensure it meets quality standards.

Machine translation can work for certain types of content. Yet, in many cases, quality lags behind human standards. Our quality assessment team helps you take advantage of lower costs without sacrificing quality. Venga has qualified linguists in over 75 languages to evaluate your machine translation.

About Venga

With software industry origins, Venga is a translation, localization, and global creative partner that specializes in helping technology companies grow globally.

Clean data sets are vital for training AI to better respond to and interact with natural human language both written and verbal. By using supervised data collection methods, Venga delivers multilingual data sets of text, voice, and image data that are suitable for test sets and training sets.

At Venga, we have an extensive pool of qualified resources and can scale to 1000+ linguists working to translate your data sets into target languages in as little as 2-4 weeks. This means that you will have quality multilingual data fast.