VIRACENT

DATA ANNOTATION | AI STRATEGY | MULTILINGUAL | CROSS-INDUSTRY

Your AI Model Is Only as Good as the Data *Behind* It.

Data annotation is not a back-office preprocessing task. It is the single most consequential investment an organisation makes in the quality of its AI — and the businesses that treat it that way are building models that actually work in production.

Viracent Delivery Practice • 8 min read • Published 2026

THE OVERLOOKED FOUNDATION

Why annotation quality is where AI programmes quietly succeed or fail

Across the AI programmes we have delivered and supported — from enterprise model training for global multinationals to focused automation projects for growing businesses — a consistent pattern emerges. The organisations that achieve the most reliable, most accurate, and most commercially useful AI models are not necessarily those with the largest compute budgets or the most sophisticated architectures. They are the ones that invest seriously in the quality and consistency of their annotated training data.

Data annotation is the process of labelling raw data — text, images, audio, video, documents — so that machine learning models can learn from it. Every named entity a language model recognises, every object a vision system detects, every intent a conversational AI interprets, traces its reliability back to the quality of the annotations that shaped the model's understanding. Annotation is not a precursor to the real work. It is the real work.

Yet in our experience, it is routinely underestimated — treated as a volume exercise to be outsourced cheaply and completed quickly, rather than a precision capability that deserves the same governance and quality discipline as the model training it enables. The consequences of that underestimation show up predictably: models that perform well in controlled test environments and disappoint in production, bias patterns that trace directly to inconsistent labelling, and costly retraining cycles that could have been prevented upstream.

> *"Every retraining cycle we have been called in to resolve has had the same root cause: annotation inconsistency that was visible in the data long before it showed up in model performance."*
>
> — JITENDER, VIRACENT AI DELIVERY PRACTICE

ANNOTATION QUALITY VS. PRODUCTION MODEL ACCURACY — ILLUSTRATIVE RELATIONSHIP

Illustrative relationship based on practitioner observations. Actual figures vary by task type, domain complexity, and model architecture.

THE TRENDS SHAPING THE FIELD

How data annotation has evolved — and where it is heading

The annotation landscape has shifted substantially in the past five years. What was once a largely manual, crowd-sourced activity — large volumes of generic data labelled by distributed workforces with minimal domain expertise — has matured into a discipline that demands specialisation, quality infrastructure, and increasingly, a hybrid of human judgment and machine assistance.

Several trends are defining the current state of the field and shaping how organisations should approach annotation as a strategic investment rather than a commodity task.

WHERE COMPLEXITY LIVES

Multilingual and handwritten data — the annotation challenges most teams underestimate

As AI deployment moves from English-language, digitally-native contexts into the full breadth of global business operations, two categories of annotation complexity emerge as particularly consequential: multilingual data and handwritten content. Both are areas where generic annotation pipelines break down, and where the quality delta between careful, specialist annotation and commodity labelling is most pronounced.

Multilingual annotation is not simply translation. A sentiment classifier trained on English text does not transfer its accuracy to Arabic, Hindi, or Mandarin by running the labels through a translation engine. Each language carries its own syntactic structures, dialectal variations, cultural context, and — critically for NLP tasks — its own set of ambiguities that require human judgment anchored in genuine linguistic competence. For organisations building AI products that will serve multilingual user bases, this is not an edge case. It is the core challenge.

Handwritten data introduces a different order of complexity. The variability in handwriting — across individuals, across languages, across document age and condition — is vast in ways that printed text annotation does not prepare teams for. Medical records, legal documents, historical archives, financial forms, logistics manifests — all of these contain handwritten content that is operationally significant and routinely excluded from AI systems because the annotation investment required has not been made. Viracent's annotation practice covers both categories directly, with language-specific annotator teams and handwriting-specialist workflows that handle the full range of script complexity.

Aa

European Languages

English, French, German, Spanish, Italian & more

عر

Arabic Scripts

Arabic, Urdu, Farsi — RTL & handwritten variants

हि

Indic Scripts

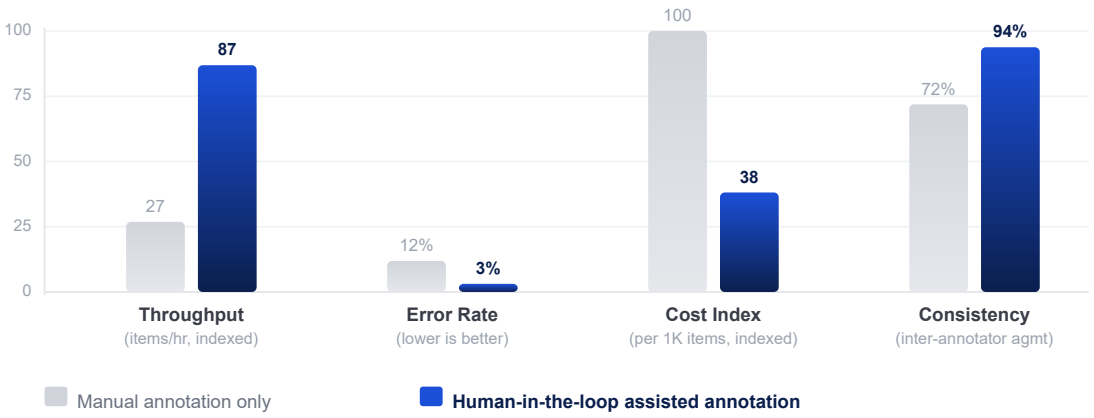
Hindi, Bengali, Tamil, Telugu & regional dialects

中文

CJK Languages

Mandarin, Japanese, Korean — typed & handwritten

MANUAL VS. HUMAN-IN-THE-LOOP ASSISTED ANNOTATION — COST & SPEED COMPARISON



Indicative figures for illustrative purposes based on practitioner observations. Results vary by task type, domain, and tool configuration. All values are indexed relative to a manual-only baseline of 100 where applicable.

WHAT VIRACENT DELIVERS

A full-service annotation capability — from strategy to production pipeline

Viracent offers data annotation as a direct service, structured around the quality and governance standards that production AI models demand. Our practice has supported multinational organisations building large-scale proprietary training datasets, as well as growing businesses establishing their annotation capability from the ground up. In both contexts, the same principles apply: annotation quality is non-negotiable, and the pipeline that produces it should be as well-engineered as the model it feeds.

VIRACENT DATA ANNOTATION SERVICE — CORE CAPABILITIES

Multilingual annotation across 30+ languages — including handwritten scripts

Our annotator network includes native-speaker specialists across European, Arabic, Indic, and CJK language families. We handle typed and handwritten content in all major scripts, with dialect-aware teams for languages where regional variation materially affects label quality. For organisations building multilingual NLP models, this is the capability that

determines whether the training data is genuinely representative of the language as it is used — not as a translation engine renders it.

Large-scale dataset construction for enterprise AI model training

We have built and annotated proprietary training datasets for multinational organisations across sectors including logistics, financial services, healthcare, and retail. These engagements range from image classification and object detection datasets for computer vision models to named entity recognition corpora for domain-specific language models. The defining characteristic of each is a quality assurance framework that produces inter-annotator agreement scores at or above the threshold required for reliable model training.

Annotation types: text, image, audio, video, and complex documents

Our annotation service covers the full modality spectrum — text classification, sentiment and intent labelling, named entity recognition, bounding box and polygon annotation for images, keypoint and landmark detection, audio transcription and speaker diarisation, video event tagging, and structured document annotation including forms, tables, and handwritten records. Complex multi-modal annotation tasks are handled by specialist teams who understand the relationship between modalities, not separate pipelines stitched together.

Quality assurance framework with measurable consistency guarantees

Every Viracent annotation engagement operates under a documented QA framework: annotation guidelines authored before labelling begins, gold-standard benchmark sets for ongoing calibration, inter-annotator agreement measurement at regular intervals, and a structured review and arbitration process for edge cases. Clients receive quality metrics with every delivery batch — not as a summary at the end, but as a continuous signal throughout the engagement.

VIRACENT DATA ANNOTATION PIPELINE — END-TO-END PROCESS

THE STRATEGIC IMPERATIVE

Proprietary annotated data as competitive advantage

The most sophisticated AI operators in every industry have arrived at the same conclusion: the models themselves are increasingly commoditised, but the proprietary datasets that train domain-specific models are not. An organisation that has invested in building a high-quality, well-annotated corpus of its own operational data — customer interactions, document archives, transaction records, sensor feeds — has an AI asset that cannot be replicated by a competitor who simply licenses the same foundation model.

This is the strategic framing that separates organisations that treat annotation as a cost to be minimised from those that treat it as a capability to be built. The former will find themselves perpetually dependent on generic models with generic performance. The latter are building a compounding advantage that grows more valuable with every annotation cycle.

WHEN TO ENGAGE A SPECIALIST ANNOTATION PARTNER

— **Your data includes non-English or handwritten content:** Generic annotation pipelines are not equipped to handle script complexity, dialectal variation, or the variability inherent in handwritten documents — specialist teams are a necessity, not a preference

— **You are building a large-scale training dataset for the first time:** The governance infrastructure — guidelines, benchmarks, QA frameworks — is as important as the labelling capacity, and harder to build internally without prior experience

— **Model performance in production is below test benchmarks:** This is almost always a data quality signal; an annotation audit is the highest-value first step

— **You need to move quickly without compromising quality:** Viracent's human-in-the-loop pipeline combines AI-assisted pre-labelling with specialist human validation to deliver throughput at quality levels that purely manual or purely automated approaches cannot match

— **You are a multinational building globally applicable AI:** Language coverage, cultural context, and regional data representation are capabilities that require deliberate investment — and a delivery partner with the annotator network to support it

Ready to build a data annotation capability that your AI can rely on?

Our team offers a complimentary Data Readiness Assessment — a focused session that evaluates your current data estate, identifies annotation gaps, and produces a clear picture of what your models need to perform reliably in production.

Book a Data Readiness Assessment →

viracent.com
ContactUs@Viracent.com