CONCEPT

Synthetic Data and AI Training Data Preparation Course: Annotation and Privacy-Safe Datasets



Synthetic Data and AI Training Data Preparation Course: Annotation and Privacy-Safe Datasets

Introduction:

Synthetic data and AI training data preparation, covering annotation and privacy-safe datasets, is a 1-week course for machine learning, labelling operations and privacy engineering teams, ending with a Training Data Preparation Plan for a case organisation. Many model projects stall because labels drift between annotators, vendors deliver inconsistent batches and sensitive records cannot be shared for training. Nominees already build or curate datasets at work; the course adds measurable labelling quality, synthetic record generation and re-identification testing, taught through case study work on real annotation samples. CoreConcept Training Center delivers this synthetic data and AI training data preparation course.

Course Objectives:

- Write annotation guidelines and label taxonomies that annotators can apply consistently to text, image and tabular records
- Measure inter-annotator agreement with Cohen's kappa, Fleiss' kappa and Krippendorff's alpha and act on low scores
- Manage labelling vendors and annotator workforces through gold sets, adjudication queues and batch acceptance criteria
- Generate synthetic tabular, text and instruction records with statistical models, GANs and LLM seed expansion
- Test candidate datasets for re-identification exposure using k-anonymity, l-diversity and differential privacy budgets
- Produce a Training Data Preparation Plan with dataset documentation ready for release to model teams

Target Audience:

- Machine learning engineering teams who assemble training and evaluation sets for production models
- Data science staff who fine-tune language models and need instruction or preference data
- Annotation and labelling operations staff who run annotator pools or supervise outsourced vendors
- Data product owners responsible for dataset scope, labelling budgets and release decisions
- Privacy engineering staff who assess disclosure exposure before records leave a controlled environment

Course Outline:

Day 1: Training Data Foundations and Current-State Dataset Review

- Supervised Learning Label Needs Across Classification and Extraction Tasks
- Training, Validation and Held-Out Test Split Design
- Instruction, Preference and Evaluation Sets for Language Model Projects
- Label Noise and Class Imbalance Effects on Model Behaviour
- Dataset Inventory Worksheet for Current-State Gap Review

Day 2: Annotation Guidelines, Taxonomies and Agreement Metrics

- Annotation Guideline Template With Edge Case Decision Rules
- Label Taxonomy Design With Hierarchy and Exclusion Criteria
- Cohen's Kappa Calculation for Paired Annotator Comparison
- Fleiss' Kappa and Krippendorff's Alpha for Annotator Pools
- Pilot Labelling Round to Calibrate Guideline Wording

Day 3: Labelling Operations, Vendors and Synthetic Record Generation

- Gold Set Seeding and Adjudication Queue Workflow
- Labelling Vendor Statement of Work and Batch Acceptance Sampling
- Active Learning Uncertainty Sampling and Query by Committee
- Statistical Model and GAN Generators for Tabular Records
- LLM Seed Instruction Expansion for Synthetic Fine-Tuning Data

Day 4: Synthetic Data Evaluation, Re-identification Risk and Safe Release

- Fidelity and Utility Scoring of Synthetic Versus Real Records
- Train-on-Synthetic Test-on-Real Benchmark for Model Transfer
- k-Anonymity and l-Diversity Checks on Quasi-Identifier Combinations
- Differential Privacy Epsilon Budget for Synthetic Data Release
- Membership Inference and Linkage Attack Testing Before Sharing

Day 5: Case Study Work and the Training Data Preparation Plan

- Retail Support Chatbot Case Guideline and Agreement Review
- Clinical Notes Case Synthetic Record and Disclosure Assessment
- Datasheets for Datasets Documentation Drafted for Case Corpus
- Data Sharing Agreement Terms Under Personal Data Protection Law
- Training Data Preparation Plan Completion and Peer Challenge

Skills You Will Gain:

- Annotation Guideline Authoring
- Label Taxonomy Design
- Inter-Annotator Agreement Analysis
- Labelling Vendor Oversight
- Active Learning Sampling
- Synthetic Record Generation
- Re-identification Risk Testing
- Dataset Documentation

Why Attend This Course:

- Deliver a Training Data Preparation Plan to the model owner and the data product lead, covering labelling, synthesis and release controls
- Decide when synthetic records can replace, augment or only test against real training data
- Avoid relabelling cycles, rejected vendor batches and disclosure incidents caused by unclear guidelines or unsafe releases
- Coach annotators and vendor leads on guideline use, agreement targets and adjudication practice

Conclusion:

Back at work, the participant hands the model owner and the data product lead a Training Data Preparation Plan for one active ML or LLM project. They use it to approve labelling budgets, set agreement targets for vendors, choose where synthetic records are acceptable and confirm that a dataset may be shared. After the first labelled batch or synthetic release, the unit should compare agreement scores, model transfer results and disclosure test findings with the plan and adjust guidelines, sampling and privacy budgets accordingly.

Frequently Asked Questions FAQ:

What should participants know before a synthetic data and AI training data preparation course?

Participants should already work with datasets for machine learning or language models and be comfortable with spreadsheets or basic Python notebooks. Bringing an anonymised annotation guideline, a sample labelled batch or a dataset description from their own project makes the case study work more useful.

How does synthetic data and AI training data preparation differ from a data governance or data readiness course?

It concentrates on producing the dataset itself: writing guidelines, measuring annotator agreement, managing labelling vendors, generating synthetic records and testing disclosure exposure. Governance and readiness courses audit policies, ownership and statutory controls across data holdings, which this course touches only where a dataset is released.

Can synthetic data fully replace real records when preparing AI training data?

Rarely. Synthetic records can supply accurate labels, rare cases and privacy-safer samples, yet models trained only on synthetic data often transfer poorly. Mixing a small real sample and testing synthetic-trained models on real held-out data shows whether the synthetic set is fit for purpose.

What does a participant take back from the synthetic data and AI training data preparation course?

A Training Data Preparation Plan for one project, with an annotation guideline, agreement targets, vendor acceptance criteria, a synthetic generation approach, disclosure test results and a datasheet. Model owners can use it directly to approve labelling and release decisions.