# AI Lifecycle Tool Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No | Stage | Technical Details | Platform-Specific Enhancements | Granular Tasks/Details | Key Tools & GPU Performance |
| 1 | Data Ingestion | Ingest structured/unstructured data from S3, APIs, Kafka, DBs | NetApp S3 provides scalable, high-speed parallel I/O | Create ETL pipelines; configure S3 buckets | Tools: AWS S3, NetApp S3, Kafka, Airbyte GPU Impact: 4.8 TB/s memory bandwidth enables rapid streaming via GPUDirect Storage |
| 2 | Data Preparation | ETL pipelines, cleansing, deduplication, annotation | Run:AI jobs for parallelized data-prep containers on OCP | Null handling; data-type normalization | Tools: Pandas, Spark, Run:AI GPU Impact: 141 GB HBM3e processes 100M rows in-memory |
| 3 | Data Labeling | Annotation of raw data for supervised learning | Managed labeling workflows on OCP with GPU-accelerated UI | Define schemas; assign tasks; review labels | Tools: Label Studio; Amazon SageMaker Ground Truth GPU Impact: UI rendering at 60 FPS for large datasets |
| 4 | Feature Engineering | Vectorization, embeddings, normalization, dimensionality reduction | GPU acceleration for embeddings & AutoML feature selection | BERT embeddings; PCA | Tools: Hugging Face; RAPIDS cuDF/cuML; PyTorch GPU Impact: 1 M tokens/sec for BERT-large (FP16) |
| 5 | Vector Database | Indexing & retrieval of high-dimensional embeddings | Sidecar deployment on OCP with GPU-accelerated ANN search | Build & maintain vector index; run similarity queries; upsert embeddings | Tools: Pinecone; Qdrant; ChromaDB; Weaviate; Milvus GPU Impact: ~10 M vectors/sec similarity search via GPU offload |
| 6 | Model Selection | Choose transformer models, CNNs, RNNs, or LLMs | Supports Hugging Face, PyTorch, TensorFlow on H200 GPUs | Load pretrained models | Tools: Hugging Face Hub; ONNX Model Hub GPU Impact: FP8 allows testing Llama3-14B on 8 GPUs |
| 7 | Model Training | Distributed training with DDP, Horovod, DeepSpeed | Run:AI orchestrates GPU scheduling across 24× H200 GPUs | Mixed precision; gradient checkpointing | Tools: PyTorch Lightning; DeepSpeed GPU Impact: 4.2× faster training for 70B models vs H100 |
| 8 | Model Evaluation | Validation metrics (accuracy, F1) on test sets | On-demand jobs via OpenShift, tracked via Run:AI dashboards | Confusion matrix; precision/recall evaluation | Tools: MLflow; Weights & Biases GPU Impact: 512 samples/sec batch inference for Llama2-70B (FP8) |
| 9 | Model Optimization | Optimize models for inference (quantization, pruning) | GPU-accelerated optimization pipelines on OCP | Apply pruning; quantize to INT8/FP16 | Tools: TensorRT; ONNX Runtime; DeepSpeed inference GPU Impact: 2–4× reduced latency and model size |
| 10 | Hyperparameter Tuning | Grid search, Bayesian optimization, early stopping | Run:AI enables multiple parallel tuning jobs | Optuna/Ray Tune jobs | Tools: Optuna; Ray Tune GPU Impact: 24-GPU parallelism cuts tuning time by 5.2× |
| 11 | Model Packaging | Convert to ONNX/TorchScript; containerize via Docker | Deploy containers on OpenShift with GPU access | Export to ONNX | Tools: ONNX; Docker (NVIDIA Container Toolkit); TensorRT GPU Impact: TensorRT boosts inference speed by 1.44× |
| 12 | Model Deployment | Batch/realtime deployment via REST, gRPC or Kafka | OpenShift CI/CD; GPU inference via Triton | Deploy Triton server; autoscaling | Tools: NVIDIA Triton; Seldon Core GPU Impact: 2.1 ms/token latency for Llama3-7B (FP8) |
| 13 | Monitoring & Feedback | Drift detection, logging, performance dashboards | NetApp logging + Run:AI reports + OCP observability | GPU metrics visualization | Tools: Prometheus; Grafana GPU Impact: 99% GPU utilization via Run:AI monitoring |
| 14 | Model Retraining | Trigger retraining on feedback/drift | Run:AI automates job queueing & scheduled retraining | Version new models | Tools: MLflow; Kubeflow GPU Impact: 38% faster retraining with FP8 |