Finalized Pipeline Strategy and Model Snapshot Policy 2025-03-31

Date: 2025-03-31

Pipeline Maturity Highlights

- Pipeline tested successfully on multi-day JSON flattening and transformation.
- All preprocessing stages validated: dtype optimization, NLP embeddings, label encodings, regex flags.
- Model-ready DataFrame fully validated and saved as .pkl and .csv.
- Test data merged, audited, and deduplicated safely corpus integrity confirmed.

Encoder Safety Policy

- Encoders are saved monthly to prevent index drift or mismatches during merges.
- Future data merges must reuse original encoders or retrain encoders over the entire corpus.
- Best practice is to rerun the full pipeline monthly to ensure vocabulary consistency.

Model Snapshot Best Practice

- Monthly snapshot includes: model_ready.pkl, text_embeddings.npz, embedding_encoders.pkl.
- Snapshots stored in folder structure: /data/processed/YYYY/MM/
- Supports traceability, reproducibility, and post-hoc debugging in production ML pipelines.

Folder Structure Strategy

- Raw data organized as: /data/raw/YYYY/MM/
- Processed data organized as: /data/processed/YYYY/MM/
- Temporary testing directory used: /data/processed/test/
- Merged corpus maintained under: /data/master/

Next Steps

- Run preprocessing on full final-March JSON scrape batch.
- Save model-ready dataset and artifacts under 2025/03.
- Validate output via CSV structure check.
- Proceed to transformer data batching + architecture config.