TrackTempo Automation Summary & Next Steps

Automation Achievements

Preprocessing Pipeline:

- Flatten raw JSON racecards with batch runner
- Clean, enrich, and embed categorical fields
- NLP vectorization for comment + spotlight fields
- Saves outputs: .csv, .pkl, text embeddings, encoders (timestamped)

Merging Pipeline:

- Merges inference dataset with race results
- Handles fuzzy joins, suffix normalization, missing result audits
- Saves: final training set, unmatched horses, missing races (all timestamped)

Training Pipeline:

- Batches and masks race data with winner labels
- Trains a RaceTransformer with checkpointing
- Loads latest encoder automatically
- Saves checkpoints by epoch to timestamped folder

Suggested Next Steps

Master CLI or Streamlit Controller:

- Enable toggles: [x] Preprocess, [x] Merge, [x] Train
- Later: [x] Evaluate, [x] Inference

Evaluation Enhancements:

- Validation splitting (holdout races or horses)
- Metrics: Accuracy, ROC-AUC, Top-1/Top-3 Hits
- Epoch logging to CSV or JSON
- Save loss/accuracy curves

Leaderboards & Auditing:

- Visualize predicted vs actual winners
- Show ranked predictions per race
- Display confusion matrix

Production Export:

- TorchScript / ONNX save for trained model
- Ready for API or frontend inference

Filesystem Refactor:

- Group: flattening, utils, batching into logical units
- Avoid 3+ level deep from x.y.z import calls
- Keep a clean tracktempo/ root

Current File Naming Convention

All pipeline outputs are timestamped:

- inference_dataset_YYYY-MM-DDTHH-MM.pkl
- model_ready_train_YYYY-MM-DDTHH-MM.pkl
- embedding_encoders_YYYY-MM-DDTHH-MM.pkl
- text_embeddings_YYYY-MM-DDTHH-MM.npz

Logs:

- unmatched.csv
- missing_races.csv