# TrackTempo Evaluation Pipeline Plan

## Evaluation Goals

- Measure model performance at race-level granularity
- Visualize prediction confidence and rank accuracy
- Enable leaderboard-style audits of horse-by-horse predictions
- Export results for inspection, visualization, and further analysis

## Planned Script: evaluate_pipeline.py

Inputs:
- --model: Path to trained model checkpoint (.pt)
- --data: Path to model_ready_test.pkl (or sliced .pkl)
- --encoders: Path to embedding_encoders.pkl

Optional Flags:
- --save_csv: Export ranked predictions to CSV
- --visualize: Render matplotlib plots

## Evaluation Logic

1. Load model and test dataset
2. Feed forward each race batch
3. Compute:
    - Softmax prediction scores
    - True vs predicted winner flag
    - True vs predicted ranks
4. Metrics:
    - Accuracy (Top-1 Hit)
    - Top-3 Hit Rate
    - Predicted Rank vs Actual Rank correlation

5. Output:

- Printed summary
- (Optional) CSV of ranked predictions
- (Optional) Matplotlib bar + scatter plots

**Future Enhancements**

- Epoch-level evaluation metrics during training
- Group-level evaluation (by course, class, or distance)
- Streamlit leaderboard interface
- Race-by-race audit tools with per-horse breakdown
- Incorporate uncertainty / margin of confidence in predictions