TrackTempo Transformer Pipeline: Full Lifecycle Overview

Date: 2025-03-30

1. Preprocessing

- Load and flatten raw JSON data.
- Clean and feature-engineer structured fields.
- Embed text fields (spotlight, comment, etc.).
- Save model-ready dataset: model_ready.pkl

2. Train/Test Split

- Split data by race_id or date into train and test sets.
- Ensure no leakage (no overlapping races).

3. Train Transformer Model

- Use training races from model_ready.pkl.
- Run transformer model training.
- Save model weights.

4. Inference (Test Set)

- Run transformer on test races.
- Get predicted win probabilities (softmax).
- Save predictions to model_preds.csv or similar.

5. Merge Predictions + Post-Race Results

- Load post-race CSV and parse with parse_postrace_fields().

- Merge predictions into model_enriched dataset.

6. Evaluation

- Filter evaluable rows (with SP and position).
- Compute:
    - Accuracy@1
    - ROI simulation
    - Value margin and overlay analysis
    - Calibration and ranking performance
- Output evaluation summary PDF and CSV