Transformer Horse Racing Project - Refresher Summary

Date: 2025-04-02 04:56

What Weve Completed

- Flattened and cleaned multi-day JSON data into model-ready .pkl files
- Validated dtypes, handled missing fields, and encoded categoricals
- Extracted NLP embeddings from 'comment' and 'spotlight' fields
- Built regex-based feature indicators for race-related phrases
- Validated March dataset and exported full audit summary

Infrastructure & Design Decisions

- Model: Encoder-only Transformer
- Field size: Only include races with >= 5 runners
- Truncation logic: Cap race size with logging/masking
- NLP embeddings: Stored (not expanded) in each row
- Edge case plan: 'mini_mlp' for races with < 5 runners
- File structure: Saved by month in /data/processed/YYYY/MM
- Current token usage: ~27,600 / 128,000 context limit

Learning & Tooling Plan

- Following free DeepLearning.AI short courses (e.g., Efficient Fine-Tuning, LLM Efficiency)
- Testing PCA downsampling of NLP vectors as optional compression strategy
- Interested in LoRA, quantization, memory-efficient attention for future improvements

Coming Up Next

- Define batch_races() utility for dynamic padding
- Group races and pad within each batch to max runner count
- Prepare float, categorical, and NLP tensors for training
- Define transformer input heads and architecture
- Begin training loop for ranking/winning prediction

Current Status

- Pre-Batching Phase: ready to build batching utilities for transformer ingestion