Daily Summary - Project Pipeline Status

Date: 2025-03-29

Today's Accomplishments

- Cleaned and flattened daily JSON data to model-ready format using `flatten_day_batch.py`.
- Converted all relevant categorical fields into embedding indices (e.g., `country_idx`, `going_idx`).
- Stored text embeddings for `comment` and `spotlight` using `SentenceTransformer` and saved as `*_vector` columns.
- Extracted over 30 domain-specific binary NLP features using regex (e.g., `mentions_ground_form`, `mentions_classy_rival`).
- Merged text embedding + regex feature extraction into `process_text_fields()` utility (modular NLP + NER pipeline).
- Saved text embeddings to `.npz` format using `save_embeddings_npz()`, and added `load_embeddings_npz()` with schema validation.
- Noted: must use `**embeddings_dict` when saving `.npz` to preserve named keys (e.g., `comment`, `spotlight`).

Pending Before Transformer Pretraining

- Clean directory naming conventions for processed and model-ready files.
- Checkpoint key feature lists (e.g., float_features, embedding_features, regex_features).
- Create unit tests for core utilities (`process_text_fields`, `extract_race_phrases`, etc).
- Final cell-level integrity validation on the model-ready CSV (data quality pass).
- Launch TRANSFORMER PRETRAINING.

Next Session Plan

Upload model-ready CSV to perform final cell-level data integrity checks before moving into transformer pretraining.