# Source: docs/FINAL_REPORT.md

# Teacher-Guided Semantic Basis Projection: A General Semantic-to-Time-Series Framework (Excitement Case Study)

## Executive Abstract

This report presents a full project narrative for a method that maps abstract semantic constructs into interpretable time series. The concrete case study is narrative excitement, but the method is intentionally formulated as a general framework. The pipeline starts from long-form text, constructs sliding-window embedding trajectories, produces LLM teacher labels under three protocols, learns a linear semantic basis projection with a one-layer perceptron, selects the best teacher protocol using trend-fidelity criteria, and then analyzes structure through unsupervised clustering.

The central hypothesis is that sentence-embedding geometry contains latent semantic directions for abstract concepts such as excitement. Instead of maximizing variance as PCA does, this work optimizes semantic alignment to teacher labels. Under the locked selection rule, CIW-5 is selected as the primary teacher protocol. The results support a practical workflow for teacher-guided, interpretable semantic signal extraction from embeddings.

## 1. Project Objective and Contribution Statement

The project objective is to establish and evaluate a reproducible method for converting abstract semantics into chunk-level time-series signals that are interpretable, measurable, and suitable for downstream analysis. This differs from conventional document-level sentiment analysis because the focus is trajectory behavior across narrative progression.

Main contribution of this stage:
1. A teacher-guided semantic basis projection formulation that is simple and explicit.
2. A variant-comparison protocol for teacher labels that prioritizes trend fidelity.
3. A clustering analysis layer over the selected signal to derive pacing archetypes and genre-linked structure.
4. A packaging workflow where every claim is mapped to reproducible artifacts.

## 2. Problem Formulation and Hypothesis

Let `x_t ∈ R^D` be the sentence embedding of chunk `t` for a book. The student model predicts semantic intensity with a single linear map:

`ŷ_t = x_t^T w + b`

where `w ∈ R^D` is the semantic basis vector and `b` is a scalar bias. Training minimizes mean squared error with L2 regularization:

`L = (1/N) Σ_t (ŷ_t - y_t)^2 + λ ||w||_2^2`

Hypothesis: if excitement is encoded in embedding geometry, a supervised linear axis should recover trend-aligned signals against teacher labels. This objective is distinct from PCA, which optimizes variance explanation without semantic supervision.

## 3. Data and Representation Pipeline

The corpus contains `20` novels and `21337` total chunks across all books. Chunk counts range from `255` to `4608` per book. Novel-level split is deterministic and leakage-safe, with `16` train novels and `4` test novels.

Split profile:

split	n_books	n_chunks	median_T	min_T	max_T

test	4	3706	981.500	471	1272
train	16	17631	717.000	255	4608

Representation route used in this report:
`Data -> Sliding-window embeddings -> Teacher pseudo-ground truth variants -> Linear semantic basis projection -> Variant selection -> Clustering -> Utility analysis`

## 4. Teacher Protocols and Pseudo-Ground-Truth Design

Teacher labels are generated by LLM judging on a 0-4 excitement scale. Labels are treated as pseudo-ground truth because they are model-derived supervision, not direct human annotation. Three variants are used:

1. `NC-1`: No-Context Chunk Teacher Labels (`label.npy`).
2. `SW-5`: Shared-Window Labels (`label_winsize_5.npy`).
3. `CIW-5`: Context-Window Independent Labels (`label_indep_winsize_5.npy`).

The variant study is essential because teacher protocol changes alter supervision smoothness, local consistency, and calibration behavior.

## 5. Student Model: Semantic Basis Extraction

The student is a one-layer perceptron over standardized embeddings. The learned vector `w` is interpreted as a semantic basis direction in embedding space rather than a black-box latent representation. This gives a clear mapping between representation and predicted signal while preserving computational simplicity.

Training configuration for CIW-5 model (from saved model artifact):

1. `seed = 42`
2. `lr = 0.01000`
3. `epochs = 200`
4. `batch_size = 4096`
5. `weight_decay = 0.000100`

## 6. Evaluation Protocol and Trend-Fidelity Criterion

Primary selection policy is trend-fidelity-first on test data with MA(5) smoothing. For each variant, ranking is determined by:

1. Lowest `MAE_MA5`
2. Then lowest raw `MAE`
3. Then lowest `RMSE`

Secondary diagnostics (`R2`, correlation) are reported for context but do not override the primary criterion.

## 7. Variant Study Results and Selection Rationale

Test-split variant diagnostics:

variant_code	rmse	mae	mae_ma5	r2	corr	rank_trend_primary	rank_raw_error	rank_corr
CIW-5	0.749	0.613	0.270	0.075	0.306	1	1	3
NC-1	1.193	1.041	0.460	0.136	0.380	2	3	2
SW-5	0.989	0.768	0.621	0.152	0.411	3	2	1

Interpretation:

1. `CIW-5` ranks first by the locked trend-fidelity rule.
2. `CIW-5` test raw metrics are `RMSE=0.749`, `MAE=0.613`.
3. `CIW-5` trend metric is `MAE_MA5=0.270`, with raw-to-smoothed drop `0.343`.
4. `selected_variant_code = CIW-5` based on deterministic ranking.

## 8. Selected Variant (CIW-5) Deep Behavior Analysis

Per-test-book CIW-5 diagnostics are summarized below. `error_gap_raw_vs_ma` indicates the improvement from raw MAE to MA(5) MAE.

book_id	title	T	mae	mae_ma	corr	error_gap_raw_vs_ma
1342	Pride and Prejudice	1272	0.588	0.260	0.263	0.327
768	Wuthering Heights	1158	0.615	0.264	0.327	0.351
113	The Secret Garden	805	0.611	0.272	0.205	0.339
16	Peter Pan	471	0.678	0.309	0.283	0.369

This table supports two conclusions:
1. Trend-level agreement is consistently better than chunk-level agreement.
2. Book-level heterogeneity remains substantial, so deployment should prioritize comparative trend profiling over absolute chunk score decisions.

## 9. Clustering on Selected Signal and Archetype Interpretation

Clustering in the final presentation is applied to CIW-5 derived trajectory features using the feature branch.

Clustering input definition in this project:
1. Feature-branch clustering is performed on a per-book feature vector extracted from the **raw CIW-5 trajectory**.
2. This feature vector also includes three MA(5)-derived summary features (`mean_ma5`, `std_ma5`, `p95_ma5`).
3. Therefore, clustering is **not** performed on CIW-5 MA(5)-only sequence values. It is performed on a mixed descriptor set dominated by raw CIW-5 statistics plus MA(5) summaries.
4. MA(5) trajectories are used for additional visualization and archetype interpretation panels.

Extracted per-book features from CIW-5 time series:
1. Length and level/distribution: `T`, `mean_y`, `std_y`, `median_y`, `iqr_y`, `min_y`, `max_y`, `p10_y`, `p90_y`, `range_y`.
2. Label composition: `prop_label_0`, `prop_label_1`, `prop_label_2`, `prop_label_3`, `prop_label_4`, `entropy_labels`.
3. Local dynamics: `mean_abs_diff`, `std_diff`, `p95_abs_diff`, `jump_ge_2_rate`, `up_rate`, `down_rate`, `flat_rate`, `lag1_autocorr`, `sign_change_rate`.
4. Position/trend structure: `corr_with_position`, `slope_position`, `mean_early`, `mean_mid`, `mean_late`.
5. Smoothed summaries (MA5): `mean_ma5`, `std_ma5`, `p95_ma5`.

Selected feature configuration: method `kmeans`, `k=5`, silhouette `0.260`, stability ARI `0.893`.

Cluster summary (feature-primary):

cluster	n_books	top_feature_1	top_feature_2	top_feature_3	representative_book	dominant_genre	dominant_genre_prop
1	5	mean_early	p90_y	mean_ma5	1260 | Jane Eyre	Adventure	0.400
2	6	prop_label_2	prop_label_0	p10_y	768 | Wuthering Heights	Adventure	0.500
3	6	std_diff	jump_ge_2_rate	mean_abs_diff	1661 | The Adventures of Sherlock Holmes	Sci-Fi	0.500
4	1	prop_label_4	mean_late	slope_position	1513 | Romeo and Juliet	Tragedy	1.000
5	2	max_y	range_y	lag1_autocorr	11 | Alice's Adventures in Wonderland	Children's Fiction	0.500

Reading guidance: feature clusters are interpreted through engineered trajectory descriptors and MA(5) member trajectories. This keeps the final presentation focused on interpretable archetypes from CIW-5 features.

## 10. Figure Explanations and Evidence
### Figure 1. Pipeline Overview
![Figure 1. Pipeline Overview](../outputs/final_report/figures/fig01_pipeline_overview.png)

- What this figure shows: The full workflow from data ingestion to semantic time-series applications.
- How to read it: Read left to right. Each box is a stage and arrows indicate dependency flow.
- Interpretation and insight: The report contribution is centered on the supervised semantic-axis extraction stage, not on unsupervised variance decomposition.

### Figure 2. Variant Comparison on Test Split
![Figure 2. Variant Comparison on Test Split](../outputs/final_report/figures/fig02_variant_comparison_test_metrics.png)

- What this figure shows: Comparative test metrics for NC-1, SW-5, and CIW-5 across raw and smoothed errors plus correlation diagnostics.
- How to read it: For RMSE/MAE/MAE_MA5 lower is better. For R2/correlation higher is better. Selection still follows trend-first ranking.
- Interpretation and insight: CIW-5 is selected because it is best on the primary trend metric while remaining competitive on raw metrics.

### Figure 3. CIW-5 Model Behavior
![Figure 3. CIW-5 Model Behavior](../outputs/final_report/figures/fig03_ciw5_model_behavior.png)

- What this figure shows: Scatter, residual, and training diagnostics associated with the selected CIW-5 student model.
- How to read it: Use scatter for calibration spread, residual histogram for bias shape, and loss curves for optimization stability.
- Interpretation and insight: The model is stable and interpretable, but chunk-level residual spread confirms that trend-level interpretation is the safer use mode.

### Figure 4. CIW-5 Test Overlays
![Figure 4. CIW-5 Test Overlays](../outputs/final_report/figures/fig04_ciw5_test_overlays_reference.png)

- What this figure shows: Overlay of teacher and student trajectories for all held-out test novels.
- How to read it: Track directional movement and pacing regions instead of exact pointwise matching.
- Interpretation and insight: The selected model preserves broad narrative dynamics on unseen novels, which justifies trend-level utility claims.

### Figure 5. Feature Cluster Map
![Figure 5. Feature Cluster Map](../outputs/final_report/figures/fig05_feature_cluster_map.png)

- What this figure shows: Feature-space map of books with selected feature-cluster assignments.
- How to read it: Each point is one book and color indicates cluster identity.
- Interpretation and insight: The map provides the geometric context for archetype interpretation in the cluster summary table.

### Figure 6. Genre Composition by Cluster
![Figure 6. Genre Composition by Cluster](../outputs/final_report/figures/fig06_cluster_genre_composition.png)

- What this figure shows: Cluster composition shown in both counts and row-normalized proportions by `genre_primary`.

- How to read it: Left panel shows absolute counts. Right panel shows within-cluster composition.
- Interpretation and insight: Genre concentration varies across clusters, supporting the claim that the extracted signal captures narratively meaningful structure.

### Figure 7. Feature Cluster Signatures and Member Trajectories

![Figure 7. Feature Cluster Signatures and Member Trajectories](../outputs/final_report/figures/fig07_cluster_signatures_and_agreement.png)

- What this figure shows: Top feature signatures by cluster together with MA(5) member-trajectory archetypes.
- How to read it: Use the signature panel to read which features distinguish each cluster, then inspect MA(5) trajectory panels for pacing shape patterns.
- Interpretation and insight: Together, these views connect feature-level semantics to observable cluster trajectory behavior in CIW-5.

### Figure 8. Variant Rank Sensitivity

![Figure 8. Variant Rank Sensitivity](../outputs/final_report/figures/fig08_variant_rank_sensitivity.png)

- What this figure shows: Rank matrix of variants under trend-first, raw-error-first, and correlation-first criteria.
- How to read it: Lower rank numbers are better. Compare rows (variants) across columns (criteria).
- Interpretation and insight: This figure makes selection logic transparent and shows how conclusions shift under alternate objectives.

### Figure 9. CIW-5 Per-Book Test Breakdown

![Figure 9. CIW-5 Per-Book Test Breakdown](../outputs/final_report/figures/fig09_ciw5_per_book_test_breakdown.png)

- What this figure shows: Book-level raw MAE and MA(5) MAE bars with correlation line for the test set.
- How to read it: Compare paired bars within each book to inspect smoothing gains and use line markers for correlation context.
- Interpretation and insight: The figure quantifies where trend-level gains are strongest and where residual uncertainty remains.

### Figure 10. Contribution and Use-Cases Map

![Figure 10. Contribution and Use-Cases Map](../outputs/final_report/figures/fig10_contribution_and_use_cases_map.png)

- What this figure shows: Conceptual mapping from method components to research contributions and practical uses.
- How to read it: Read each row as input component -> methodological contribution -> usage pathway.
- Interpretation and insight: The project is positioned as a reusable semantic-to-time-series framework, with excitement as the demonstrated task.

### Figure 11. Feature Cluster Member Trajectories (MA5)

![Figure 11. Feature Cluster Member Trajectories (MA5)](../outputs/final_report/figures/fig11_feature_cluster_member_trajectories_ma5.png)

- What this figure shows: A dedicated high-resolution view of MA(5) member trajectories for each feature cluster.
- How to read it: Each subplot represents one feature cluster with thin lines for member books and a bold centroid trajectory.
- Interpretation and insight: This figure provides direct visual evidence of pacing archetypes that define the final feature-cluster interpretation.

## 11. Utility, Generalization, and Deployment Scenarios

The method is useful when a team needs interpretable, chunk-level semantic trajectories from high-dimensional embeddings and cannot afford heavy black-box sequence models. Since the student is linear and trained with explicit supervision, each run produces a transparent semantic basis that is fast to apply to new data.

Potential use scenarios:
1. Narrative pacing analytics for editorial workflow support.
2. Cross-book comparative profiling for literary or media research.
3. Teacher-student distillation pipeline for other abstract constructs (for example suspense, urgency, or emotional intensity).
4. Lightweight semantic monitoring where interpretability and reproducibility are mandatory.

## 12. Limitations and Threats to Validity
1. Teacher labels are pseudo-ground truth and can contain systematic LLM bias.
2. Corpus size is small (`20` novels), limiting external validity.
3. The student is linear, so nonlinear semantic structure may be underfit.
4. Cluster structure is sensitive to feature design and sample size.
5. Correlation and R2 behavior can diverge from trend-error objectives, so objective choice must be explicit.

## 13. Reproducibility and Artifact Guide
Generated in stage `10_final_teacher_guided_semantic_basis_report.ipynb` with deterministic configuration (`SEED=42`, `MA_WINDOW=5`).

Core evidence artifacts:
1. `../outputs/final_report/tables/variant_selection_summary.csv`
2. `../outputs/final_report/tables/variant_selection_diagnostics.csv`
3. `../outputs/final_report/tables/dataset_profile_for_report.csv`
4. `../outputs/final_report/tables/ciw5_per_book_deepdive.csv`
5. `../outputs/final_report/tables/cluster_summary_for_report.csv`
6. `../outputs/final_report/tables/key_results_registry.csv`
7. `../outputs/final_report/tables/method_claims_checklist.csv`
8. `../outputs/final_report/tables/report_integrity_checks.csv`

Related appendix: `docs/OTHER_EXPERIMENTS.md` documents Twist Signal and PCA tracks as secondary experiments, intentionally separated from the main claim path.

## 14. Conclusion
This project demonstrates a concrete path for converting abstract semantics into interpretable time series by combining embedding trajectories, LLM teacher supervision, and linear semantic basis extraction. In this case study, CIW-5 is the most suitable teacher protocol under trend-first selection. The resulting signal supports meaningful clustering and practical downstream interpretation. More broadly, the workflow provides a reusable pattern for teacher-guided semantic projection where transparency, reproducibility, and analytical utility are first-class goals.

## Claim Provenance
- Core registry: `../outputs/final_report/tables/key_results_registry.csv`
- Claim checklist: `../outputs/final_report/tables/method_claims_checklist.csv`
- Report integrity: `../outputs/final_report/tables/report_integrity_checks.csv`

# Source: docs/OTHER_EXPERIMENTS.md

# Other Experiments: Twist Signal and PCA Tracks

## Scope
This document summarizes additional experiments completed in the project that ar

e not part of the primary final claim. The primary claim is centered on Teacher-Guided Semantic Basis Projection. The experiments below remain valuable, but they are intentionally separated to keep narrative focus clear.

## Twist Signal Track

Twist Signal experiments model local novelty dynamics from embedding trajectories using `s_t` and acceleration `a_t`. This branch supports exploratory narrative-change analysis and peak detection.

Key outputs:

- `../outputs/features.csv`
- `../outputs/clusters_kmeans.csv`
- `../outputs/clusters_hier.csv`
- `../outputs/dtw_distance_k7.npy`
- `../outputs/eda/`

## PCA Component Track

PCA experiments analyze unsupervised axes of variance and their temporal behavior across books. This is useful for structural diagnostics and exploratory component interpretation.

Key outputs:

- `../outputs/pca/global_pca_fit.npz`
- `../outputs/pca/global_pca_fit_meta.json`
- `../outputs/pca/global_pca_variance_summary.csv`
- `../outputs/pca_analysis/`

## Why This Is Secondary in the Final Narrative

The final narrative aims to evaluate supervised extraction of one specific semantic basis from embeddings using teacher labels. Twist Signal and PCA branches address different questions. They are retained as supporting evidence of broad project exploration and as future integration candidates, but they are not used as primary evidence for the teacher-guided semantic basis claim.

# Source: docs/PIPELINE.md

# Pipeline Documentation

## Pipeline Stages

1. **Download and Clean (`00_download_and_clean.ipynb`)**

- Downloads Gutenberg plain-text books from configured URLs.
- Removes Gutenberg boilerplate.
- Saves cleaned raw text to `data/raw/{raw_filename}`.
- Writes metadata and catalog files.

2. **Chunk and Embed (`01_chunk_and_embed.ipynb`)**

- Splits each book into overlapping word windows.
- Encodes chunks with Sentence-Transformers.
- Saves chunk index + embeddings + run metadata under `data/processed/{processed_dir}/`.

3. **Transform and Cluster (`02_transform_and_cluster.ipynb`)**

- Fits PCA and saves low-dimensional trajectories.
- Persists global PCA fit artifacts (`outputs/pca/global_pca_fit.*`) for downstream reproducible analysis.
- Computes **Twist Signal** (`s_t`) and acceleration (`a_t`) for multiple `k` values.
- Detects top peaks and builds story-level features.
- Runs feature-based clustering and DTW-based similarity clustering.

4. **EDA and Visualization (`03_eda_and_visualization.ipynb`)**
- Loads outputs and metadata.
- Builds statistical + interactive views.
- Produces exploratory and deeper interpretation notes.
- Exports EDA figures/tables and insights.

5. **All-Novel Stacked Panels (`04_novel_stacked_twist_signal.ipynb`)**
- Loads per-book `signals_k{5,7,11}.npz` and `peaks_k{5,7,11}.json`.
- Generates one 3-row stacked figure per novel (`s_t` + `a_t` for each `k`).
- Exports grouped tables and a consolidated per-novel interpretation markdown.

6. **LLM Judge Overlay Prep (Prompt Package)**
- Builds long-context Gemini prompt payloads per novel from full raw text and chunk index lines.
- Produces strict JSON peak labels for overlay comparison (judge labels, not absolute ground truth).
- Validates output schema and spacing/range constraints before downstream plotting.

7. **LLM Judge Signal Analysis (`05_llm_judge_signal_analysis.ipynb`)**
- Focuses on `k=7` signal interpretation with smoothing + robust normalization + composite event score.
- Compares LLM peaks against existing pipeline peaks and processed event peaks.
- Exports analysis figures/tables and run-specific insight markdown under `outputs/llm_judge/analysis/`.

8. **PCA Component Insights (`06_pca_component_insights.ipynb`)**
- Loads persisted global PCA fit artifacts from `outputs/pca/`.
- Produces corpus-level PCA diagnostics and per-book PCA trajectory metrics.
- Links PCA dynamics to Twist Signal behavior for `k=5,7,11`.
- Runs permutation/bootstrapped robustness checks and exports integrity tables.
- Writes figures/tables/insight markdown under `outputs/pca_analysis/`.

9. **Excitement Linear Projection (`07_excitement_linear_projection.ipynb`)**
- Loads per-book embeddings (`embeddings.npy`), LLM excitement labels (`label.npy`), and metadata.
- Uses deterministic novel-level split with seed `42` (`16` train novels, `4` test novels).
- Trains a 1-layer linear perceptron (`y_hat = X @ W + b`) with MSE optimization.
- Exports diagnostics, raw-vs-smoothed overlay figures, metrics tables, and learned weights.
- Writes outputs under `outputs/excitement_linear/`.

10. **Excitement Label Variant Analysis (`08_excitement_label_variant_analysis.ipynb`)**
- Compares three label variants: `label.npy`, `label_winsize_5.npy`, and `label_indep_winsize_5.npy`.
- Reuses split manifest from `outputs/excitement_linear/tables/split_manifest.csv` for direct comparability.
- Trains one linear model per variant with matched optimization settings.
- Exports cross-variant agreement/distribution diagnostics plus an `indep_winsize_5` deep dive.
- Writes figures/tables/model artifacts under `outputs/excitement_variant_analysis/`.

11. **Indep Excitement Clustering (`09_indep_excitement_clustering.ipynb`)**
- Focuses on `label_indep_winsize_5.npy` for all books (unsupervised, no train/test split).
- Extracts per-book interpretable trajectory features (distribution, volatility, transition, temporal-shape, and MA-derived stats).

- Runs feature-branch clustering (`KMeans`, `Agglomerative Ward`) with `k=2..6` model sweep and stability checks.
- Runs DTW-branch clustering on resampled trajectories (`L=200`) for trajectory-shape validation.
- Computes cross-method agreement (ARI/NMI + contingency table) and exports representatives/profile summaries.
- Adds advanced diagnostics: elbow (`k=1..10`), feature/DTW k-sweep quality plots, genre-by-cluster summaries, feature-signature and contingency heatmaps.
- Adds legend integrity checks for feature-cluster visualizations.
- Writes outputs under `outputs/excitement_indep_clustering/`, including `insights.md` and `cluster_report.md`.

12. **Final Teacher-Guided Semantic Basis Report Packaging (`10_final_teacher_guided_semantic_basis_report.ipynb`)**
- Reuses stage-08 variant metrics and stage-09 clustering outputs as source-of-truth evidence.
- Builds final claim-safe support tables under `outputs/final_report/tables/`, including dataset profile, variant-rank diagnostics, CIW-5 per-book deep-dive metrics, and claim-evidence checklist.
- Generates curated storytelling figures under `outputs/final_report/figures/`, including variant-rank sensitivity, per-test-book CIW-5 breakdown, and contribution/use-case framing diagrams.
- Writes final narrative documents:
  - `docs/FINAL_REPORT.md` (primary narrative)
  - `docs/OTHER_EXPERIMENTS.md` (Twist/PCA appendix, explicitly secondary)
- Runs packaging integrity checks (source existence, table/figure completeness, image-link validity, claim-checklist completeness, selected-variant consistency, split consistency, and no em-dash policy in generated docs).

## Twist Signal Definition
Given chunk embeddings `e_t` and context window size `k`:

- Context mean: `context_mean[t] = mean(e_{t-k}, ..., e_{t-1})` using available prefix when `t < k`.
- Twist Signal: `s_t = 1 - cosine(e_t, context_mean[t])`.
- Acceleration: `a_t = |s_t - s_{t-1}|` with `a_0 = 0`.

Note: this method was formerly referenced as “Option B”; project docs now use **Twist Signal**.

## Parameters and Defaults
Current defaults in pipeline notebooks:
- Chunking:
  - `window_words=300`
  - `stride_words=100`
- Embedding:
  - `batch_size=64`
  - current run model observed in artifact indexes: `sentence-transformers/all-mpnet-base-v2`
- Twist Signal:
  - `k_values=[5, 7, 11]`
  - primary reporting at `k=7`
- PCA:
  - dimensions `2` and `5`
- Peaks:
  - top peaks `top_K=3`
  - minimum separation `3` chunks
- Clustering:
  - default clusters `n_clusters=4`
- DTW:
  - signal resample length `L=200`
- Excitement linear projection:

  - split: novel-level `16/4` with `SEED=42`
  - optimizer defaults: `EPOCHS=200`, `BATCH_SIZE=4096`, `LR=1e-2`, `WEIGHT_DECAY=1e-4`
  - presentation smoothing: runtime-configured `MA_WINDOW` (current run: `9`)
- Excitement variant analysis:
  - same split policy as `07` (manifest reuse)
  - trains one linear model per label variant
  - presentation smoothing: runtime-configured `MA_WINDOW` (current run: `5`)
- Indep excitement clustering:
  - label source: `label_indep_winsize_5.npy`
  - smoothing-derived features: runtime `MA_WINDOW` (current clustering run: `5`)
  - model sweep: `k=2..6`
  - DTW resample length: `L=200`
- Final report packaging:
  - method naming:
    - `NC-1` = No-Context Chunk Teacher Labels
    - `SW-5` = Shared-Window Labels
    - `CIW-5` = Context-Window Independent Labels
  - variant ranking rule: trend-fidelity first on test split (`mae_ma5`, then `mae`, then `rmse`)
  - selected variant for current run: `CIW-5`
  - report smoothing policy: `MA_WINDOW=5`

## Artifact Flow
File-level flow:
1. `data/raw/{raw_filename}.txt`
2. `data/processed/{processed_dir}/chunks.jsonl`
3. `data/processed/{processed_dir}/embeddings.npy`
4. `data/processed/{processed_dir}/label.npy`
5. `data/processed/{processed_dir}/label_winsize_5.npy`
6. `data/processed/{processed_dir}/label_indep_winsize_5.npy`
7. `data/processed/{processed_dir}/index.json`
8. `data/processed/{processed_dir}/pca_d2.npy`, `pca_d5.npy`
9. `outputs/pca/global_pca_fit.npz`
10. `outputs/pca/global_pca_fit_meta.json`
11. `outputs/pca/global_pca_variance_summary.csv`
12. `data/processed/{processed_dir}/signals_k{K}.npz`
13. `data/processed/{processed_dir}/peaks_k{K}.json`
14. `outputs/features.csv`
15. `outputs/clusters_kmeans.csv`, `outputs/clusters_hier.csv`
16. `outputs/dtw_distance_k7.npy`
17. `outputs/eda/*` from EDA notebook
18. `outputs/eda/novel_stacks/figures/*.png`
19. `outputs/eda/novel_stacks/tables/*.csv`
20. `docs/NOVEL_STACKED_OUTPUT_INTERPRETATION.md`
21. `outputs/llm_judge/prompts/*.json` (generated by helper script)
22. `outputs/llm_judge/analysis/figures/*.png`
23. `outputs/llm_judge/analysis/tables/*.csv`
24. `outputs/llm_judge/analysis/insights_k7.md`
25. `outputs/pca_analysis/figures/*.png`
26. `outputs/pca_analysis/tables/*.csv`
27. `outputs/pca_analysis/insights.md`
28. `outputs/excitement_linear/figures/*.png`
29. `outputs/excitement_linear/tables/*.csv`
30. `outputs/excitement_linear/model/linear_weights.npz`
31. `outputs/excitement_linear/interpretation.md`
32. `outputs/excitement_variant_analysis/figures/*.png`
33. `outputs/excitement_variant_analysis/tables/*.csv`
34. `outputs/excitement_variant_analysis/model/linear_weights_*.npz`
35. `outputs/excitement_variant_analysis/insights.md`
36. `outputs/excitement_indep_clustering/tables/*.csv`

37. `outputs/excitement_indep_clustering/figures/*.png`
38. `outputs/excitement_indep_clustering/insights.md`
39. `outputs/excitement_indep_clustering/cluster_report.md`
40. `outputs/final_report/figures/fig01_pipeline_overview.png`
41. `outputs/final_report/figures/fig02_variant_comparison_test_metrics.png`
42. `outputs/final_report/figures/fig03_ciw5_model_behavior.png`
43. `outputs/final_report/figures/fig04_ciw5_test_overlays_reference.png`
44. `outputs/final_report/figures/fig05_feature_cluster_map.png`
45. `outputs/final_report/figures/fig06_cluster_genre_composition.png`
46. `outputs/final_report/figures/fig07_cluster_signatures_and_agreement.png`
47. `outputs/final_report/figures/fig08_variant_rank_sensitivity.png`
48. `outputs/final_report/figures/fig09_ciw5_per_book_test_breakdown.png`
49. `outputs/final_report/figures/fig10_contribution_and_use_cases_map.png`
50. `outputs/final_report/figures/fig11_feature_cluster_member_trajectories_ma5.png`
51. `outputs/final_report/tables/dataset_profile_for_report.csv`
52. `outputs/final_report/tables/variant_selection_summary.csv`
53. `outputs/final_report/tables/variant_selection_diagnostics.csv`
54. `outputs/final_report/tables/ciw5_per_book_deepdive.csv`
55. `outputs/final_report/tables/key_results_registry.csv`
56. `outputs/final_report/tables/method_claims_checklist.csv`
57. `outputs/final_report/tables/cluster_summary_for_report.csv`
58. `outputs/final_report/tables/report_integrity_checks.csv`
59. `docs/FINAL_REPORT.md`
60. `docs/OTHER_EXPERIMENTS.md`

Directory keying:
- `processed_dir` is the canonical per-book key and is based on abbreviated title slug.

## Reproducibility and Caching
- Seeds are fixed in notebooks (`SEED=42`) for sampling/clustering consistency.
- Embedding stage supports cache reuse if existing artifacts match expected parameters.
- PCA fit metadata is persisted to `outputs/pca/global_pca_fit_meta.json` for downstream reproducibility checks.
- Metadata file links each `book_id` to `raw_filename` and `processed_dir`.
- If legacy numeric processed folders exist, migration/fallback logic can still resolve them.

## Failure Modes and Recovery
Common issues and recovery steps:
- **Missing raw file**:
  - Re-run `00_download_and_clean.ipynb`.
  - Verify `raw_filename` and `raw_path` in `data/metadata.csv`.
- **Missing or stale embedding cache**:
  - Re-run `01_chunk_and_embed.ipynb`.
  - Set recompute flag if parameters changed.
- **Mismatch between metadata and processed folders**:
  - Confirm `processed_dir` exists under `data/processed/`.
- **DTW shape/symmetry problems**:
  - Re-run `02_transform_and_cluster.ipynb` and verify generated matrix diagnostics.
- **Missing global PCA fit artifacts (`outputs/pca/global_pca_fit.*`)**:
  - Re-run `02_transform_and_cluster.ipynb`.
  - Confirm `outputs/pca/global_pca_variance_summary.csv` exists before running `06_pca_component_insights.ipynb`.
- **Missing or malformed `label.npy` for excitement projection**:
  - Confirm `data/processed/{processed_dir}/label.npy` exists for every book.
  - Ensure shape is one of `(T,)`, `(1,T)`, `(T,1)` and values are in `[0,4]`.
- **Embedding/label length mismatch for excitement projection**:
  - Verify `len(label) == embeddings.shape[0]` for each book before running `07_

excitement_linear_projection.ipynb`.
- **Missing `presentation_mae.csv` when interpretation is generated**:
  - Re-run `07_excitement_linear_projection.ipynb` and confirm outputs under `outputs/excitement_linear/tables/`.
- **Missing variant label files for stage 08**:
  - Confirm `label.npy`, `label_winsize_5.npy`, and `label_indep_winsize_5.npy` exist for every `processed_dir`.
- **Variant label shape/range errors**:
  - Ensure each variant label file has shape `(T,)`, `(1,T)`, or `(T,1)` and values within `[0,4]`.
- **Variant split mismatch**:
  - Re-check `outputs/excitement_linear/tables/split_manifest.csv` integrity (`16` train, `4` test, no overlap).
- **Missing stage 08 outputs**:
  - Re-run `08_excitement_label_variant_analysis.ipynb` and verify all artifacts under `outputs/excitement_variant_analysis/`.
- **Missing `label_indep_winsize_5.npy` for stage 09**:
  - Ensure each `data/processed/{processed_dir}/` has `label_indep_winsize_5.npy`.
  - Check shape compatibility `(T,)`, `(1,T)`, `(T,1)` and integer-like values in `[0,4]`.
- **Stage 09 alignment mismatch (`len(label) != embeddings.shape[0]`)**:
  - Rebuild or correct the problematic label file for the affected `processed_dir`.
  - Re-run `09_indep_excitement_clustering.ipynb`.
- **Stage 09 output incompleteness**:
  - Re-run `09_indep_excitement_clustering.ipynb` and inspect `outputs/excitement_indep_clustering/tables/integrity_checks.csv`.
- **Stage 09 legend mismatch (feature scatter legend not matching cluster count)**:
  - Inspect `outputs/excitement_indep_clustering/tables/figure_legend_checks.csv`.
  - Re-run `09_indep_excitement_clustering.ipynb` and verify `feature_pca_scatter_feature_clusters.png` legend entries.
- **Stage 09 report image-link mismatch**:
  - Verify all files under `outputs/excitement_indep_clustering/figures/` exist.
  - Re-run `09_indep_excitement_clustering.ipynb` and confirm `cluster_report_embedded_figures_exist` passes in integrity checks.
- **Missing source artifacts for stage 10 packaging**:
  - Ensure stage-08 and stage-09 tables exist before running `10_final_teacher_guided_semantic_basis_report.ipynb`.
  - Required source roots:
    - `outputs/excitement_variant_analysis/tables/`
    - `outputs/excitement_indep_clustering/tables/`
- **Stage 10 selected variant mismatch**:
  - Inspect `outputs/final_report/tables/variant_selection_summary.csv`.
  - Confirm ranking is computed from test split with trend-fidelity priority (`mae_ma5`, then `mae`, then `rmse`).
  - Re-run `10_final_teacher_guided_semantic_basis_report.ipynb` if mismatch persists.
- **Stage 10 report integrity check failure**:
  - Inspect `outputs/final_report/tables/report_integrity_checks.csv`.
  - Verify missing figures/docs paths and rerun stage 10 after repairing source artifacts.
- **Stage 10 claim-evidence mapping failure**:
  - Inspect `outputs/final_report/tables/method_claims_checklist.csv`.
  - Resolve rows with `status != mapped` by fixing metric-key references or source file paths in the stage-10 notebook generator.
- **Stage 10 image-link resolution failure**:
  - Inspect `docs/FINAL_REPORT.md` and verify every embedded `../outputs/final_report/figures/*.png` path exists.
  - Re-run stage 10 and confirm `embedded_image_links_exist` passes in `outputs/

final_report/tables/report_integrity_checks.csv`.

## Extending the Pipeline
Recommended extension points:
- Add new signal variants while preserving `signals_k{K}.npz` compatibility.
- Add cluster methods and write separate output files instead of overwriting existing schemas.
- Add model-comparison runs by tagging outputs with model metadata.
- Add report-generation notebook/scripts that consume `outputs/` and `outputs/eda/`.
- Add LLM-judge overlays using `prompts/llm_judge/` and `tools/llm_judge/` contracts.

# Source: docs/DATA_DICTIONARY.md

# Data Dictionary

## Core Keys and Joins
Primary keys used across outputs:
- `id` / `pg_id` in `data/metadata.csv`: Gutenberg book identifier.
- `book_id` in output tables: same numeric identifier as `id`.
- `processed_dir`: canonical per-book folder key under `data/processed/`.

Join rules:
- `outputs/features.csv.book_id` joins to `data/metadata.csv.id`.
- `outputs/clusters_*.csv.book_id` joins to `data/metadata.csv.id`.

## `data/metadata.csv`
Purpose: master catalog + run metadata per book.

Current columns:
- `id` (int): canonical book id used by pipeline.
- `pg_id` (int): Gutenberg id (same as `id` in current run).
- `title` (str)
- `author` (str)
- `first_publication_year` (int)
- `origin_country` (str)
- `original_language` (str)
- `format` (str)
- `genre_primary` (str)
- `genre_secondary` (JSON-string list)
- `short_tags` (JSON-string list)
- `recognizability_rank` (int)
- `genre_clarity_rank` (int)
- `twist_peak_rank` (int)
- `twist_peak_reason` (str)
- `notes` (str)
- `ebook_page_url` (str)
- `plain_text_utf8_url` (str)
- `raw_filename` (str)
- `raw_path` (str)
- `processed_dir` (str)
- `processed_path` (str)
- `length` (int): word count after cleaning
- `char_length` (int)
- `source_url` (str)
- `status` (str)
- `citations` (JSON-string list)

## Per-Book Processed Artifacts

Location pattern:

- `data/processed/{processed_dir}/...`

### `chunks.jsonl`

One JSON object per chunk:

- `chunk_index` (int)
- `start_word` (int)
- `end_word` (int)
- `text_preview` (str)

### `embeddings.npy`

- Shape: `(T, D)`
- Dtype: `float32`
- `T`: chunk count, `D`: embedding dimension

### `label.npy`

Purpose: LLM excitement label per chunk used by `07_excitement_linear_projection.ipynb`.

Contract:

- Accepted input shapes: `(T,)`, `(1, T)`, `(T, 1)`; normalized to `(T,)` in notebook loaders.
- Dtype: numeric (`int` or float-castable).
- Expected value range: integer labels in `[0, 4]`.
- Alignment rule: `len(label) == embeddings.shape[0]` for each book.

### `label_winsize_5.npy`

Purpose: LLM excitement labels produced from 5-chunk grouped prompts where each 5-chunk block is assigned one shared score.

Contract:

- Accepted input shapes: `(T,)`, `(1, T)`, `(T, 1)`; normalized to `(T,)` in notebook loaders.
- Dtype: numeric (`int` or float-castable).
- Expected value range: integer labels in `[0, 4]`.
- Alignment rule: `len(label_winsize_5) == embeddings.shape[0]`.
- Block behavior expectation: labels are constant inside each contiguous 5-chunk block.

### `label_indep_winsize_5.npy`

Purpose: LLM excitement labels produced from 5-chunk grouped prompts with independent per-chunk scoring within each group.

Contract:

- Accepted input shapes: `(T,)`, `(1, T)`, `(T, 1)`; normalized to `(T,)` in notebook loaders.
- Dtype: numeric (`int` or float-castable).
- Expected value range: integer labels in `[0, 4]`.
- Alignment rule: `len(label_indep_winsize_5) == embeddings.shape[0]`.

### `index.json`

Typical fields:

- `book_id` (int)
- `processed_dir` (str)
- `T` (int)
- `D` (int)
- `window_words` (int)
- `stride_words` (int)
- `model_name` (str)
- `batch_size` (int)
- `dtype` (str)

- `created_at` (str ISO timestamp)

### `signals_k{K}.npz`
NPZ keys:
- `s`: Twist Signal array, shape `(T,)`, dtype `float32`
- `a`: acceleration array, shape `(T,)`, dtype `float32`

### `peaks_k{K}.json`
Fields:
- `book_id` (int)
- `k` (int)
- `top_K` (int)
- `separation` (int)
- `peak_indices` (list[int])
- `peak_positions_norm` (list[float])
- `signal` (str, currently `a_t`)

### PCA Files
- `pca_d2.npy`: shape `(T, 2)`
- `pca_d5.npy`: shape `(T, 5)`

## Global PCA Artifacts (`outputs/pca/`)

### `global_pca_fit.npz`
Purpose: persisted fitted PCA model arrays used to reproduce per-book projection checks and downstream PCA analysis.

NPZ keys and shapes:
- `components`: shape `(5, D)`, dtype `float32`
- `explained_variance`: shape `(5,)`, dtype `float32`
- `explained_variance_ratio`: shape `(5,)`, dtype `float32`
- `singular_values`: shape `(5,)`, dtype `float32`
- `mean`: shape `(D,)`, dtype `float32`

### `global_pca_fit_meta.json`
Purpose: reproducibility metadata for the global PCA fit.

Fields:
- `seed` (int)
- `n_components` (int)
- `svd_solver` (str)
- `fit_rows_used` (int)
- `fit_rows_total` (int)
- `embedding_dim` (int)
- `model_name` (str or null)
- `created_at` (str ISO timestamp)

### `global_pca_variance_summary.csv`
Columns:
- `pc` (str: `PC1`..`PC5`)
- `explained_variance_ratio` (float)
- `cumulative_explained_variance_ratio` (float; monotonic non-decreasing)

## `outputs/features.csv`
Purpose: story-level feature matrix per `(book_id, k)` enriched with metadata.

Core feature columns:
- `book_id`, `k`, `T`
- `mean_s`, `std_s`, `max_s`
- `mean_a`, `std_a`, `max_a`
- `num_peaks`
- `peak_pos_1`, `peak_pos_2`, `peak_pos_3`

- `auc_proxy_mean_s`

Also includes metadata columns such as:
- `title`, `author`, `genre_primary`, `format`, `origin_country`, rank fields, tags, citations, `processed_dir`, etc.

Expected cardinality:
- Rows = `num_books * len(k_values)`
- Current run: `20 * 3 = 60`

## `outputs/clusters_kmeans.csv`
Purpose: feature-based KMeans labels.

Columns:
- `book_id`
- `k`
- `n_clusters`
- `cluster`
- metadata context fields (`title`, `author`, `genre_primary`, `format`, `processed_dir`, etc.)

Expected cardinality:
- One row per `(book_id, k)`
- Current run: `60` rows

## `outputs/clusters_hier.csv`
Purpose: hierarchical cluster labels from two modes.

Columns:
- `book_id`
- `k`
- `mode` (`feature_ward` or `dtw_average`)
- `n_clusters`
- `cluster`
- metadata context fields (`title`, `author`, `genre_primary`, `format`, `processed_dir`, etc.)

Expected cardinality:
- `feature_ward`: one row per `(book_id, k)`
- `dtw_average`: one row per book for primary `k`
- Current run: `60 + 20 = 80` rows

## `outputs/dtw_distance_k7.npy`
Purpose: pairwise DTW distance matrix on resampled Twist Signal (`k=7`).

Contract:
- Shape `(N, N)` where `N = number of books`
- Symmetric
- Diagonal approximately zero
- Current run shape: `(20, 20)`

## Excitement Linear Projection Outputs (`outputs/excitement_linear/`)

### `tables/split_manifest.csv`
Purpose: deterministic novel-level split assignment used for training/evaluation.

Columns:
- `book_id` (int)
- `title` (str)
- `processed_dir` (str)
- `T` (int)

- `split` (`train` or `test`)

Contract:
- Expected split cardinality for current setup: `16` train novels and `4` test novels.
- No overlap between train/test book ids.

### `tables/global_metrics.csv`
Purpose: global regression metrics aggregated by split.

Columns:
- `split` (`train` or `test`)
- `n_samples` (int)
- `n_novels` (int)
- `mse` (float)
- `rmse` (float)
- `mae` (float)
- `r2` (float)

### `tables/per_novel_metrics.csv`
Purpose: per-novel regression metrics from the learned linear projection.

Columns:
- `book_id` (int)
- `title` (str)
- `processed_dir` (str)
- `split` (`train` or `test`)
- `T` (int)
- `mse` (float)
- `rmse` (float)
- `mae` (float)
- `r2` (float; may be `NaN` for constant-label edge cases)
- `mae_ma` (float): moving-average MAE using current `MA_WINDOW`.

### `tables/presentation_mae.csv`
Purpose: presentation-friendly train/test MAE comparison between raw and smoothed trajectories.

Columns:
- `split` (`train` or `test`)
- `ma_window` (int): smoothing window used in current run (`MA_WINDOW` from notebook).
- `mae_raw` (float)
- `mae_moving_average` (float)
- `n_samples` (int)

### `tables/integrity_checks.csv`
Purpose: run-level and per-book integrity checks for alignment, split integrity, model shapes, and output completeness.

Columns:
- `check` (str)
- `expected` (str or numeric)
- `actual` (str or numeric)
- `pass` (bool)

### `tables/figure_support_stats.csv`
Purpose: per-book supporting statistics used when interpreting overlay/scatter/residual figures.

Columns:
- `book_id` (int)

- `split` (`train` or `test`)
- `T` (int)
- `y_true_mean`, `y_true_std` (float)
- `y_pred_mean`, `y_pred_std` (float)
- `pred_min`, `pred_max` (float)
- `res_mean`, `res_std` (float)
- `res_p05`, `res_p95` (float)
- `corr_true_pred` (float)
- `title` (str)

### `model/linear_weights.npz`
Purpose: persisted learned linear transformation from embedding space to scalar excitement prediction.

NPZ keys:
- `W`: shape `(768, 1)`, dtype `float32`
- `b`: shape `(1,)`, dtype `float32`
- `x_mean`: shape `(768,)`, dtype `float32`
- `x_std`: shape `(768,)`, dtype `float32`
- `seed`: scalar array (`int32`)
- `lr`: scalar array (`float32`)
- `epochs`: scalar array (`int32`)
- `batch_size`: scalar array (`int32`)
- `weight_decay`: scalar array (`float32`)

### `figures/*.png`
Purpose: diagnostics and per-novel overlays for model fit and trend alignment.

Canonical outputs:
- `labels_all_20_novels_grid.png`
- `labels_all_20_novels_overlay_normpos.png`
- `train_loss_curve.png`
- `prediction_scatter_train_test.png`
- `residual_hist_train_test.png`
- `mae_raw_vs_moving_average.png`
- `novel_overlay_test_{book_id}.png` (4 files, test novels)
- `novel_overlay_train_{book_id}.png` (2 files, selected train novels)

### `interpretation.md`
Purpose: narrative interpretation of diagnostics and overlays with explicit verdict on chunk-level vs trend-level usability.

Evidence linkage:
- References figures and tables under the same `outputs/excitement_linear/` namespace.
- Uses `presentation_mae.csv.ma_window` to document current smoothing window (runtime-configurable via `MA_WINDOW`).

## Excitement Variant Analysis Outputs (`outputs/excitement_variant_analysis/`)

### `tables/integrity_checks.csv`
Purpose: consolidated validation for stage-08 inputs and outputs.

Checks include:
- variant label file existence
- shape normalization compatibility
- value range `[0,4]` and integer-like checks
- `len(label_variant) == T` alignment
- split integrity reuse checks
- output completeness and overlay counts
- MA-window contract check (`ma_window == 5` for current run)

Columns:
- `check` (str)
- `expected` (str or numeric)
- `actual` (str or numeric)
- `pass` (bool)

### `tables/split_manifest_used.csv`
Purpose: split manifest copy used by stage 08 (reused from stage 07 output).

Columns:
- `book_id` (int)
- `title` (str)
- `processed_dir` (str)
- `T` (int)
- `split` (`train` or `test`)

### `tables/label_distribution_by_variant.csv`
Purpose: label frequency summaries for `base`, `winsize_5`, and `indep_winsize_5`.

Columns:
- `variant` (`base`, `winsize_5`, `indep_winsize_5`)
- `label` (`0..4`)
- `count` (int)
- `proportion` (float)
- `scope` (`global` or `book`)
- `book_id` (int or null for global rows)
- `processed_dir` (str or null for global rows)
- `title` (str or null for global rows)

### `tables/variant_pairwise_agreement_global.csv`
Purpose: global chunk-level pairwise agreement between label variants.

Columns:
- `variant_a` (str)
- `variant_b` (str)
- `mae` (float)
- `exact_match` (float)
- `corr` (float)
- `n_samples` (int)

### `tables/variant_pairwise_agreement_per_book.csv`
Purpose: per-book pairwise agreement between label variants.

Columns:
- `book_id` (int)
- `processed_dir` (str)
- `variant_a` (str)
- `variant_b` (str)
- `mae` (float)
- `exact_match` (float)
- `corr` (float)
- `T` (int)

### `tables/model_global_metrics_by_variant.csv`
Purpose: split-level regression metrics for each variant-specific linear model.

Columns:
- `variant` (str)
- `split` (`train` or `test`)
- `n_samples` (int)
- `n_novels` (int)

- `mse` (float)
- `rmse` (float)
- `mae` (float)
- `r2` (float)
- `corr` (float)
- `ma_window` (int)
- `mae_ma` (float)

### `tables/model_per_novel_metrics_by_variant.csv`
Purpose: per-book regression metrics for each variant-specific linear model.

Columns:
- `variant` (str)
- `book_id` (int)
- `title` (str)
- `processed_dir` (str)
- `split` (`train` or `test`)
- `T` (int)
- `mse` (float)
- `rmse` (float)
- `mae` (float)
- `r2` (float)
- `corr` (float)
- `mae_ma` (float)

### `tables/indep_winsize_5_support_stats.csv`
Purpose: per-book support stats for `indep_winsize_5` interpretation figures.

Columns:
- `book_id` (int)
- `split` (`train` or `test`)
- `T` (int)
- `y_true_mean`, `y_true_std` (float)
- `y_pred_mean`, `y_pred_std` (float)
- `pred_min`, `pred_max` (float)
- `res_mean`, `res_std` (float)
- `res_p05`, `res_p95` (float)
- `corr_true_pred` (float)
- `title` (str)

### `model/linear_weights_base.npz`
### `model/linear_weights_winsize_5.npz`
### `model/linear_weights_indep_winsize_5.npz`
Purpose: per-variant learned linear projection weights.

NPZ keys:
- `W`: shape `(768, 1)`, dtype `float32`
- `b`: shape `(1,)`, dtype `float32`
- `x_mean`: shape `(768,)`, dtype `float32`
- `x_std`: shape `(768,)`, dtype `float32`
- `seed`: scalar array (`int32`)
- `lr`: scalar array (`float32`)
- `epochs`: scalar array (`int32`)
- `batch_size`: scalar array (`int32`)
- `weight_decay`: scalar array (`float32`)
- `variant`: scalar string array

### `figures/*.png`
Purpose: cross-variant diagnostics and indep-focused deep-dive visuals.

Canonical outputs:
- `labels_grid_base.png`

- `labels_grid_winsize_5.png`
- `labels_grid_indep_winsize_5.png`
- `label_overlay_normpos_by_variant.png`
- `label_distribution_by_variant.png`
- `variant_pairwise_agreement_bar.png`
- `train_loss_curves_by_variant.png`
- `model_metric_comparison_by_variant.png`
- `indep_prediction_scatter_train_test.png`
- `indep_residual_hist_train_test.png`
- `indep_mae_raw_vs_moving_average.png`
- `indep_novel_overlay_test_{book_id}.png` (4 files)
- `indep_novel_overlay_train_{book_id}.png` (2 files)

### `insights.md`
Purpose: narrative summary of variant comparison and explicit `indep_winsize_5` verdict.

Contract notes:
- Uses MA(W) wording with current run `W=5`.
- Must include “use now / avoid now” guidance and next-step acceptance criteria.

## Indep Excitement Clustering Outputs (`outputs/excitement_indep_clustering/`)

### `tables/indep_book_features.csv`
Purpose: one-row-per-book feature matrix extracted from `label_indep_winsize_5.npy`.

Columns:
- `book_id`, `title`, `processed_dir`
- `T`
- `mean_y`, `std_y`, `median_y`, `iqr_y`, `min_y`, `max_y`
- `p10_y`, `p90_y`, `range_y`
- `prop_label_0`, `prop_label_1`, `prop_label_2`, `prop_label_3`, `prop_label_4`
- `entropy_labels`
- `mean_abs_diff`, `std_diff`, `p95_abs_diff`, `jump_ge_2_rate`
- `up_rate`, `down_rate`, `flat_rate`
- `lag1_autocorr`, `sign_change_rate`
- `corr_with_position`, `slope_position`
- `mean_early`, `mean_mid`, `mean_late`
- `mean_ma5`, `std_ma5`, `p95_ma5`

Contract:
- Exactly one row per book (`20` rows in current corpus).
- No NaN/inf in numeric columns.

### `tables/indep_book_features_zscore.csv`
Purpose: z-scored version of clustering features (across books).

Columns:
- Same schema as `indep_book_features.csv` with numeric columns standardized.

### `tables/cluster_quality_by_method.csv`
Purpose: model-selection diagnostics across clustering branches and candidate `k`.

Columns:
- `branch` (`feature` or `dtw`)
- `method` (`kmeans`, `ward`, or `average`)
- `k` (int, tested `2..6`)
- `silhouette` (float)
- `davies_bouldin` (float; `NaN` for DTW branch where not used)
- `calinski_harabasz` (float; `NaN` for DTW branch where not used)

- `kmeans_stability_ari` (float; populated for KMeans rows)

### `tables/cluster_assignments_feature.csv`
Purpose: final cluster assignment per book from selected feature-branch solution.

Columns:
- `book_id`, `title`, `processed_dir`
- `method`
- `k`
- `cluster`

### `tables/cluster_assignments_dtw.csv`
Purpose: final cluster assignment per book from selected DTW-branch solution.

Columns:
- `book_id`, `title`, `processed_dir`
- `method`
- `k`
- `cluster`

### `tables/cluster_profile_summary.csv`
Purpose: cluster-level feature profiling with effect-size style deltas versus corpus.

Columns:
- `branch` (`feature` or `dtw`)
- `cluster` (int)
- `cluster_size` (int)
- `feature` (str)
- `mean_raw` (float)
- `median_raw` (float)
- `mean_z` (float)
- `delta_z_vs_corpus` (float)
- `abs_delta_z` (float)
- `rank_abs_delta` (int)

### `tables/cluster_representatives.csv`
Purpose: representative-book selection per cluster for interpretation.

Columns:
- `branch` (`feature` or `dtw`)
- `cluster` (int)
- `role` (`centroid_medoid`, `cluster_medoid`, `high_volatility`, `low_volatility`)
- `book_id`, `title`, `processed_dir`
- `score_name` (str)
- `score_value` (float)

### `tables/cluster_method_agreement.csv`
Purpose: cross-method agreement metrics plus contingency table export.

Columns:
- `row_type` (`metric` or `contingency`)
- `metric` (`ari`, `nmi`, or null for contingency rows)
- `value` (float; populated for metric rows)
- `feature_cluster` (int or null)
- `dtw_cluster` (int or null)
- `count` (int; populated for contingency rows)

### `tables/kmeans_elbow_curve.csv`
Purpose: KMeans elbow diagnostics computed on feature z-space.

Columns:
- `k` (int, `1..10`)
- `inertia` (float)
- `delta_inertia` (float; previous-k inertia drop)
- `pct_drop_from_prev` (float)

Contract:
- `k` coverage must include all integers from `1` to `10`.
- `inertia` should be monotonic non-increasing across increasing `k`.

### `tables/genre_by_feature_cluster_counts.csv`
Purpose: feature-cluster by genre contingency table (counts).

Schema:
- Index column: `cluster` (int)
- Data columns: one column per `genre_primary` value.
- Cell values: integer counts of books.

### `tables/genre_by_feature_cluster_proportions.csv`
Purpose: row-normalized feature-cluster by genre table.

Schema:
- Index column: `cluster` (int)
- Data columns: one column per `genre_primary` value.
- Cell values: proportions in `[0,1]`; each row sums to `1.0` (within numeric tolerance).

### `tables/feature_cluster_signature_top_features.csv`
Purpose: selected top signature features used for feature-cluster interpretation/heatmap.

Columns:
- `branch` (`feature`)
- `cluster` (int)
- `cluster_size` (int)
- `feature` (str)
- `mean_raw` (float)
- `median_raw` (float)
- `mean_z` (float)
- `delta_z_vs_corpus` (float)
- `abs_delta_z` (float)
- `rank_abs_delta` (int; within-cluster rank)
- `global_feature_rank` (int; rank by max `abs_delta_z` across feature clusters)

### `tables/figure_legend_checks.csv`
Purpose: programmatic verification that figure legends match expected cluster entries.

Columns:
- `figure` (str)
- `expected_entries` (int)
- `actual_entries` (int)
- `pass` (bool)
- `labels` (str; serialized legend labels)

### `tables/integrity_checks.csv`
Purpose: validation summary for stage 09.

Checks include:
- label existence/shape/range/integer-like checks for all books
- label/embedding length alignment

- feature and clustering output integrity
- selected `k` range checks
- elbow monotonicity and k-coverage checks
- feature-scatter legend entry checks
- genre table consistency checks
- cluster-report embedded figure existence checks
- output completeness checks
- MA-window contract check (`MA_WINDOW=5` for current run)

Columns:
- `check` (str)
- `expected` (str or numeric)
- `actual` (str or numeric)
- `pass` (bool)

### `figures/*.png`
Purpose: visual diagnostics for features, cluster structure, and trajectory archetypes.

Canonical outputs:
- `feature_correlation_heatmap.png`
- `feature_pca_scatter_feature_clusters.png`
- `feature_elbow_kmeans_inertia.png`
- `feature_k_sweep_quality_metrics.png`
- `dtw_k_sweep_silhouette.png`
- `genre_by_feature_cluster_counts.png`
- `genre_by_feature_cluster_proportions.png`
- `cluster_feature_signature_heatmap_top12.png`
- `cluster_method_contingency_heatmap.png`
- `feature_cluster_member_trajectories_ma5.png`
- `cluster_centroid_trajectories_raw.png`
- `cluster_centroid_trajectories_ma5.png`
- `dtw_distance_heatmap.png`
- `dtw_dendrogram.png`
- `cluster_size_comparison.png`

### `insights.md`
Purpose: indep-focused clustering narrative with selected feature/DTW solutions, agreement reading, archetype summaries, and next-step guidance.

Contract notes:
- Uses MA(W) wording with current run `W=5`.
- Must include cluster-selection rationale and representative-book interpretation.

### `cluster_report.md`
Purpose: extended clustering interpretation report with embedded figure gallery and evidence-linked conclusions.

Contract notes:
- Uses relative image links to `figures/*.png`.
- Must include sections for verdict, diagnostics, feature-cluster interpretation, genre composition, feature-vs-DTW agreement, and practical guidance.

## Final Report Packaging Outputs (`outputs/final_report/`)

### `tables/variant_selection_summary.csv`
Purpose: deterministic variant-ranking table for final-report selection using trend-fidelity-first criteria.

Columns:
- `variant_code` (`NC-1`, `SW-5`, `CIW-5`)

- `variant_name` (str)
- `split` (`train` or `test`)
- `rmse` (float)
- `mae` (float)
- `mae_ma5` (float)
- `r2` (float)
- `corr` (float)
- `rank_trend_primary` (int)

Selection contract:
- Primary ranking is applied on `split=test`.
- Ranking keys are, in order:
  - `mae_ma5` (ascending)
  - `mae` (ascending)
  - `rmse` (ascending)
- `r2` and `corr` are diagnostics and do not override ranking.

### `tables/dataset_profile_for_report.csv`
Purpose: explicit corpus profile used in final report narrative and split reproducibility checks.

Columns:
- `book_id` (int)
- `title` (str)
- `processed_dir` (str)
- `genre_primary` (str; `Unknown` fallback if missing)
- `T` (int; chunk count)
- `split` (`train` or `test`)

Contract:
- One row per book in split manifest.
- Current run expectation: `20` rows with split counts `16` train and `4` test.

### `tables/variant_selection_diagnostics.csv`
Purpose: expanded ranking diagnostics showing variant behavior under multiple selection criteria.

Columns:
- `variant_code` (`NC-1`, `SW-5`, `CIW-5`)
- `split` (`train` or `test`)
- `rmse` (float)
- `mae` (float)
- `mae_ma5` (float)
- `r2` (float)
- `corr` (float)
- `rank_trend_primary` (int; ranking by `mae_ma5`, `mae`, `rmse`)
- `rank_raw_error` (int; ranking by `mae`, `rmse`, `mae_ma5`)
- `rank_corr` (int; ranking by `corr`, `r2`, `mae_ma5`)

Contract:
- Used to explain rank sensitivity across evaluation objectives.
- Current selection rule remains `rank_trend_primary` on `split=test`.

### `tables/ciw5_per_book_deepdive.csv`
Purpose: per-book diagnostics table for selected variant (`CIW-5`) with raw-vs-smoothed error gap.

Columns:
- `book_id` (int)
- `title` (str)
- `split` (`train` or `test`)
- `T` (int)

- `mse` (float)
- `rmse` (float)
- `mae` (float)
- `mae_ma` (float; MA(5)-smoothed MAE)
- `corr` (float)
- `r2` (float)
- `error_gap_raw_vs_ma` (float; `mae - mae_ma`)

Contract:
- Contains all books for the selected variant.
- Test-subset rows are used for report figure `fig09_ciw5_per_book_test_breakdown.png`.

### `tables/key_results_registry.csv`
Purpose: claim registry so numeric report statements are source-traceable and reproducible.

Columns:
- `metric_key` (str)
- `value` (float or string)
- `source_file` (str path)
- `source_row_filter` (str; row selection used to recover value)
- `notes` (str)

Contract:
- Every core numeric claim used in `docs/FINAL_REPORT.md` should have a corresponding row.

### `tables/method_claims_checklist.csv`
Purpose: claim-to-evidence checklist that maps report claims to metric keys or source files.

Columns:
- `claim_id` (str; stable claim identifier, e.g., `CLM01`)
- `claim_text` (str)
- `metric_key_or_source` (str; semicolon-delimited metric keys and/or `path:` references)
- `status` (`mapped` or `missing`)

Contract:
- `status` must be `mapped` for all rows in a passing report run.
- Referenced metric keys should exist in `tables/key_results_registry.csv`.

### `tables/cluster_summary_for_report.csv`
Purpose: compact cluster summary table for final report narrative.

Columns:
- `cluster` (int)
- `n_books` (int)
- `top_feature_1` (str)
- `top_feature_2` (str)
- `top_feature_3` (str)
- `representative_book` (str)
- `dominant_genre` (str)
- `dominant_genre_prop` (float)

Source dependencies:
- `outputs/excitement_indep_clustering/tables/cluster_profile_summary.csv`
- `outputs/excitement_indep_clustering/tables/cluster_representatives.csv`
- `outputs/excitement_indep_clustering/tables/genre_by_feature_cluster_proportions.csv`

### `tables/report_integrity_checks.csv`
Purpose: packaging-stage integrity checks for final report generation.

Columns:
- `check` (str)
- `expected` (str or numeric)
- `actual` (str or numeric)
- `pass` (bool)

Checks include:
- source artifact existence
- selected variant consistency (`CIW-5` expected for current run)
- required final-report table existence/count checks
- curated figure existence/count
- final document existence
- embedded image-link validity in `docs/FINAL_REPORT.md`
- claim-checklist completeness (`method_claims_checklist.csv`)
- split consistency checks (`16/4` for current run)
- no em-dash policy checks in generated docs

### `figures/fig01_pipeline_overview.png`
Purpose: process diagram from data through embeddings, teacher labels, linear projection, clustering, and applications.

### `figures/fig02_variant_comparison_test_metrics.png`
Purpose: test-split variant comparison panel (RMSE, MAE, MAE(MA5), R2, correlation) with trend-first selection context.

### `figures/fig03_ciw5_model_behavior.png`
Purpose: CIW-5 diagnostic composite (scatter, residual distribution, optimization behavior).

### `figures/fig04_ciw5_test_overlays_reference.png`
Purpose: montage/reference panel for CIW-5 test-novel trajectory overlays.

### `figures/fig05_feature_cluster_map.png`
Purpose: feature-space cluster geometry view used in final narrative.

### `figures/fig06_cluster_genre_composition.png`
Purpose: genre-by-cluster composition view (counts and proportions).

### `figures/fig07_cluster_signatures_and_agreement.png`
Purpose: feature-cluster signature heatmap paired with MA(5) member-trajectory archetype visualization (feature branch focus).

### `figures/fig08_variant_rank_sensitivity.png`
Purpose: rank-sensitivity matrix comparing variant ranking under trend-first, raw-error-first, and correlation-first criteria.

### `figures/fig09_ciw5_per_book_test_breakdown.png`
Purpose: selected-variant (`CIW-5`) per-test-book diagnostics panel showing raw MAE, MA(5) MAE, and correlation.

### `figures/fig10_contribution_and_use_cases_map.png`
Purpose: conceptual map from pipeline components to methodological contribution and downstream use cases.

### `figures/fig11_feature_cluster_member_trajectories_ma5.png`
Purpose: dedicated high-resolution panel of feature-cluster member trajectories using MA(5) smoothing.

## Final Narrative Documents

### `docs/FINAL_REPORT.md`
Purpose: main final narrative centered on Teacher-Guided Semantic Basis Projection.

Contract notes:
- Must frame LLM labels as pseudo-ground truth.
- Must use naming convention `NC-1`, `SW-5`, `CIW-5`.
- Must explain figure evidence before interpretation for each major figure block.
- Must keep Twist/PCA details secondary and refer readers to `docs/OTHER_EXPERIMENTS.md`.

### `docs/OTHER_EXPERIMENTS.md`
Purpose: narrative appendix for secondary experiment tracks (Twist Signal and PCA), separated from the primary final claim.

Contract notes:
- Should summarize scope, key outputs, and what was learned.
- Should explicitly state that these tracks are not the central evidence for the final teacher-guided projection claim.

## PCA Analysis Outputs (`outputs/pca_analysis/`)

### `tables/book_component_stats.csv`
Purpose: per-book PCA trajectory summary statistics.

Core columns:
- `book_id`, `processed_dir`, `title`, `genre_primary`, `T`
- `mean_pc1..mean_pc5`
- `std_pc1..std_pc5`
- `corr_pc1_position..corr_pc5_position`
- `sign_change_rate_pc1..sign_change_rate_pc5`
- `sign_change_rate_mean`
- `mean_speed`, `p95_speed`, `speed_std`

### `tables/book_component_signal_assoc.csv`
Purpose: per-book association metrics between PCA trajectory speed and Twist Signal for each `k`.

Columns:
- `book_id`, `processed_dir`, `title`, `genre_primary`
- `k`
- `T`
- `corr_speed_s`
- `corr_speed_a`
- `mean_speed`
- `p95_speed`

Expected cardinality:
- One row per `(book_id, k)` for books with valid `signals_k{K}.npz`
- With complete artifacts: `num_books * 3`

### `tables/component_exemplar_chunks.csv`
Purpose: high-scoring positive/negative chunk exemplars per PCA component for interpretation.

Columns:
- `book_id`, `processed_dir`, `title`, `genre_primary`
- `chunk_index`
- `pc` (`PC1`..`PC5`)
- `direction` (`positive` or `negative`)

- `score`
- `text_preview`

Selection constraints:
- Minimum chunk spacing (same book/component/direction): `>= 5`
- Maximum selected chunks per book/component/direction: `3`
- Target selected rows per component+direction: up to `15`

### `tables/component_genre_association.csv`
Purpose: genre-level PCA component aggregation with effect-size style normalization.

Columns:
- `genre_primary`
- `pc`
- `genre_mean`
- `genre_book_count`
- `corpus_mean`
- `corpus_std`
- `delta_vs_corpus_mean`
- `effect_size_vs_corpus`

### `tables/temporal_trend_stats.csv`
Purpose: per-book temporal trend significance for `corr(PC, normalized_position)`.

Columns:
- `book_id`, `processed_dir`, `title`, `genre_primary`
- `pc`
- `corr_pc_position`
- `perm_pvalue` (two-sided permutation p-value)
- `perm_qvalue` (Benjamini-Hochberg adjusted within each component family)

### `tables/corpus_assoc_bootstrap.csv`
Purpose: bootstrap confidence intervals for corpus-level medians of PCA-speed to Twist Signal correlations.

Columns:
- `k`
- `metric` (`corr_speed_s` or `corr_speed_a`)
- `median`
- `ci_lower`
- `ci_upper`
- `n_books`

### `tables/projection_consistency_checks.csv`
Purpose: sampled consistency checks between recomputed PCA projection and stored `pca_d5.npy`.

Columns:
- `book_id`, `processed_dir`
- `checked` (bool)
- `max_abs_diff`
- `mean_abs_diff`
- `detail`

### `tables/pca_integrity_checks.csv`
Purpose: run-level validation summary covering shape, association, statistical, exemplar, and end-to-end checks.

Columns:
- `check` (str)

- `passed` (bool)
- `detail` (str)

### `tables/book_artifact_integrity.csv`
Purpose: per-book missing/mismatch artifact logging during PCA analysis loading.

Columns:
- `book_id`, `processed_dir`
- `issue`
- `severity` (`warning` or `error`)

### `tables/book_signal_assoc_issues.csv`
Purpose: per-book and per-`k` skip logging for signal-association calculations.

Columns:
- `book_id`, `processed_dir`, `k`
- `issue`
- `severity`

### `figures/*.png`
Representative files:
- `pca_variance_diagnostics.png`
- `component_score_distributions.png`
- `component_pairwise_by_genre.png`
- `temporal_trend_summary.png`
- `bootstrap_assoc_summary.png`
- `book_deep_dive_speed_signal_k7.png`

### `insights.md`
Purpose: narrative summary of component interpretation, book-level highlights, and sensitivity/caveats.

## Type and Parsing Notes
Fields serialized as JSON-like strings in CSV:
- `genre_secondary`
- `short_tags`
- `citations`

Recommended parsing strategy:
- Try `json.loads` first.
- Fallback to `ast.literal_eval` if needed.
- Normalize lists to either list objects or comma-separated strings depending on task.

## Validation Rules
Minimum checks for every run:
1. `metadata.id` is unique and non-null.
2. `metadata.processed_dir` exists for all rows.
3. Every `processed_dir` contains required per-book artifact files.
4. `features.csv` has complete `(book_id, k)` coverage.
5. No NaN/inf in numeric Twist Signal feature columns.
6. `dtw_distance_k7.npy` passes square/symmetric/zero-diagonal checks.

# Source: README.md

# Story Trajectory Analysis

## Project Overview
This project builds story trajectories from public-domain books and analyzes nar

rative change patterns over time.

Pipeline goals:
- Download and clean Gutenberg texts.
- Convert each story into chunk-level embedding time series.
- Compute a **Twist Signal** and derived acceleration signal.
- Build story-level features and cluster books by trajectory behavior.
- Support EDA, visualization, and insight generation on top of saved artifacts.

## What Is the Twist Signal?
For each chunk embedding `e_t`, the pipeline computes a local context embedding from previous chunks, then scores novelty relative to that context.

Definitions:
- `context_mean[t]`: mean embedding over recent previous chunks (window size `k`).
- `s_t = 1 - cosine(e_t, context_mean[t])`.
- `a_t = |s_t - s_{t-1}|`.

Interpretation:
- `s_t` (Twist Signal): how different the current chunk is from recent context.
- `a_t`: how sharply novelty changes from one step to the next.

## Current Dataset Snapshot
Current run status (from local artifacts):
- Corpus size: **20 books**.
- Feature rows: **60** (`20 books * 3 k-values`).
- `k` values: `[5, 7, 11]`.
- Current embedding model (from `data/processed/*/index.json`): `sentence-transformers/all-mpnet-base-v2`.
- DTW matrix: `outputs/dtw_distance_k7.npy` shape `(20, 20)`.

Primary metadata source:
- `data/metadata.csv`
- `data/book_catalog.json`

## Repository Structure
- `00_download_and_clean.ipynb`: Gutenberg download + cleaning + metadata write.
- `01_chunk_and_embed.ipynb`: chunking + embedding + per-book processed artifacts.
- `02_transform_and_cluster.ipynb`: Twist Signal, PCA, clustering, DTW.
- `03_eda_and_visualization.ipynb`: EDA, visualization, and insight extraction.
- `04_novel_stacked_twist_signal.ipynb`: all-novel stacked Twist Signal panels (`k=5,7,11`) + consolidated per-novel interpretation export.
- `05_llm_judge_signal_analysis.ipynb`: `k=7` LLM-judge alignment analysis with processed event score, overlays, and insight exports.
- `06_pca_component_insights.ipynb`: corpus/book-level PCA component interpretation, robustness checks, and PCA-signal linkage outputs.
- `07_excitement_linear_projection.ipynb`: LLM excitement label visualization and linear embedding-to-excitement projection diagnostics.
- `08_excitement_label_variant_analysis.ipynb`: compares 3 excitement label variants (`base`, `winsize_5`, `indep_winsize_5`) and deep-dives on `indep_winsize_5`.
- `09_indep_excitement_clustering.ipynb`: unsupervised clustering on `label_indep_winsize_5.npy` using interpretable features plus DTW trajectory-shape validation.
- `10_final_teacher_guided_semantic_basis_report.ipynb`: reproducible final-report build stage that packages curated figures/tables and writes final narrative docs.
- `prompts/llm_judge/`: Gemini LLM-judge prompt templates + output schema.
- `tools/llm_judge/`: helper scripts to build per-book prompt payloads and validate JSON outputs.

- `data/raw/`: cleaned book text files (abbreviated title filenames).
- `data/processed/{processed_dir}/`: per-book chunk, embedding, label, signal, peak, PCA artifacts.
- `outputs/`: global outputs (`features.csv`, cluster CSVs, DTW matrix, summary).
- `outputs/eda/`: EDA figures, tables, and insight narrative.
- `outputs/eda/novel_stacks/`: grouped 20-book stacked figures, supporting tables, and validation outputs.
- `outputs/excitement_linear/`: figures, metrics tables, model weights, and interpretation for the linear excitement projection workflow.
- `outputs/excitement_variant_analysis/`: variant comparison tables/figures, per-variant linear weights, and indep-focused insights.
- `outputs/excitement_indep_clustering/`: indep-focused clustering tables/figures, compact insights, and extended report with embedded figures.
- `outputs/final_report/`: curated final-report figures, support tables for claim traceability, and report integrity checks.
- `docs/`: pipeline, schema dictionary, EDA planning, output interpretation, and Gemini judge prompt docs.

## Pipeline Run Order
Run notebooks in this order:
1. `00_download_and_clean.ipynb`
2. `01_chunk_and_embed.ipynb`
3. `02_transform_and_cluster.ipynb`
4. `03_eda_and_visualization.ipynb`
5. `04_novel_stacked_twist_signal.ipynb`
6. `05_llm_judge_signal_analysis.ipynb`
7. `06_pca_component_insights.ipynb` (requires `02` PCA artifacts)
8. `07_excitement_linear_projection.ipynb` (requires per-book `embeddings.npy` and `label.npy`)
9. `08_excitement_label_variant_analysis.ipynb` (reuses split from `outputs/excitement_linear/tables/split_manifest.csv`)
10. `09_indep_excitement_clustering.ipynb` (requires `label_indep_winsize_5.npy` for all books; unsupervised on all 20 novels; `MA_WINDOW=5` for smoothing-derived features)
11. `10_final_teacher_guided_semantic_basis_report.ipynb` (packages final report artifacts; reads outputs from stages `08` and `09`)

## Key Outputs for EDA/Insight Work
Core files:
- `outputs/features.csv`: one row per `(book_id, k)` with Twist Signal-derived metrics + metadata.
- `outputs/clusters_kmeans.csv`: feature-based KMeans labels.
- `outputs/clusters_hier.csv`: hierarchical labels (`feature_ward` and `dtw_average`).
- `outputs/dtw_distance_k7.npy`: pairwise DTW distances on resampled `s_t` for `k=7`.
- `outputs/pca/global_pca_fit.npz`: persisted fitted PCA arrays (`components`, `mean`, EVR arrays).
- `outputs/pca/global_pca_fit_meta.json`: PCA fit metadata (`seed`, rows used, model name, timestamps).
- `outputs/pca/global_pca_variance_summary.csv`: EVR + cumulative EVR table (PC1..PC5).

Per-book files:
- `data/processed/{processed_dir}/label.npy`
- `data/processed/{processed_dir}/label_winsize_5.npy`
- `data/processed/{processed_dir}/label_indep_winsize_5.npy`
- `data/processed/{processed_dir}/signals_k{K}.npz`
- `data/processed/{processed_dir}/peaks_k{K}.json`
- `data/processed/{processed_dir}/pca_d2.npy`
- `data/processed/{processed_dir}/pca_d5.npy`

Advanced per-novel stacked outputs:
- `outputs/eda/novel_stacks/figures/novel_{book_id}_{processed_dir}_stacked_k5_k7_k11.png`
- `outputs/eda/novel_stacks/tables/novel_stacked_stats.csv`
- `outputs/eda/novel_stacks/tables/novel_stacked_manifest.csv`
- `outputs/eda/novel_stacks/tables/novel_stacked_highlights.csv`
- `docs/NOVEL_STACKED_OUTPUT_INTERPRETATION.md`
- `docs/GEMINI_LLM_JUDGE_PROMPT_PACKAGE.md`
- `docs/LLM_JUDGE_SIGNAL_ANALYSIS.md`
- `outputs/llm_judge/analysis/insights_k7.md`

PCA component insight outputs:
- `outputs/pca_analysis/tables/book_component_stats.csv`
- `outputs/pca_analysis/tables/book_component_signal_assoc.csv`
- `outputs/pca_analysis/tables/component_exemplar_chunks.csv`
- `outputs/pca_analysis/tables/component_genre_association.csv`
- `outputs/pca_analysis/tables/temporal_trend_stats.csv`
- `outputs/pca_analysis/tables/corpus_assoc_bootstrap.csv`
- `outputs/pca_analysis/tables/projection_consistency_checks.csv`
- `outputs/pca_analysis/tables/pca_integrity_checks.csv`
- `outputs/pca_analysis/figures/*.png`
- `outputs/pca_analysis/insights.md`

Excitement linear projection outputs:
- `outputs/excitement_linear/figures/*.png`
- `outputs/excitement_linear/tables/split_manifest.csv`
- `outputs/excitement_linear/tables/global_metrics.csv`
- `outputs/excitement_linear/tables/per_novel_metrics.csv`
- `outputs/excitement_linear/tables/presentation_mae.csv`
- `outputs/excitement_linear/tables/integrity_checks.csv`
- `outputs/excitement_linear/tables/figure_support_stats.csv`
- `outputs/excitement_linear/model/linear_weights.npz`
- `outputs/excitement_linear/interpretation.md`

Excitement variant analysis outputs:
- `outputs/excitement_variant_analysis/figures/*.png`
- `outputs/excitement_variant_analysis/tables/integrity_checks.csv`
- `outputs/excitement_variant_analysis/tables/split_manifest_used.csv`
- `outputs/excitement_variant_analysis/tables/label_distribution_by_variant.csv`
- `outputs/excitement_variant_analysis/tables/variant_pairwise_agreement_global.csv`
- `outputs/excitement_variant_analysis/tables/variant_pairwise_agreement_per_book.csv`
- `outputs/excitement_variant_analysis/tables/model_global_metrics_by_variant.csv`
- `outputs/excitement_variant_analysis/tables/model_per_novel_metrics_by_variant.csv`
- `outputs/excitement_variant_analysis/tables/indep_winsize_5_support_stats.csv`
- `outputs/excitement_variant_analysis/model/linear_weights_base.npz`
- `outputs/excitement_variant_analysis/model/linear_weights_winsize_5.npz`
- `outputs/excitement_variant_analysis/model/linear_weights_indep_winsize_5.npz`
- `outputs/excitement_variant_analysis/insights.md`

Indep excitement clustering outputs:
- `outputs/excitement_indep_clustering/tables/indep_book_features.csv`
- `outputs/excitement_indep_clustering/tables/indep_book_features_zscore.csv`
- `outputs/excitement_indep_clustering/tables/cluster_quality_by_method.csv`
- `outputs/excitement_indep_clustering/tables/cluster_assignments_feature.csv`
- `outputs/excitement_indep_clustering/tables/cluster_assignments_dtw.csv`
- `outputs/excitement_indep_clustering/tables/cluster_profile_summary.csv`
- `outputs/excitement_indep_clustering/tables/cluster_representatives.csv`

- `outputs/excitement_indep_clustering/tables/cluster_method_agreement.csv`
- `outputs/excitement_indep_clustering/tables/kmeans_elbow_curve.csv`
- `outputs/excitement_indep_clustering/tables/genre_by_feature_cluster_counts.csv`
- `outputs/excitement_indep_clustering/tables/genre_by_feature_cluster_proportions.csv`
- `outputs/excitement_indep_clustering/tables/feature_cluster_signature_top_features.csv`
- `outputs/excitement_indep_clustering/tables/figure_legend_checks.csv`
- `outputs/excitement_indep_clustering/tables/integrity_checks.csv`
- `outputs/excitement_indep_clustering/figures/*.png` (including elbow, k-sweep, genre, contingency, and cluster-member panels)
- `outputs/excitement_indep_clustering/insights.md`
- `outputs/excitement_indep_clustering/cluster_report.md`

Final report packaging outputs:
- `outputs/final_report/figures/fig01_pipeline_overview.png`
- `outputs/final_report/figures/fig02_variant_comparison_test_metrics.png`
- `outputs/final_report/figures/fig03_ciw5_model_behavior.png`
- `outputs/final_report/figures/fig04_ciw5_test_overlays_reference.png`
- `outputs/final_report/figures/fig05_feature_cluster_map.png`
- `outputs/final_report/figures/fig06_cluster_genre_composition.png`
- `outputs/final_report/figures/fig07_cluster_signatures_and_agreement.png`
- `outputs/final_report/figures/fig08_variant_rank_sensitivity.png`
- `outputs/final_report/figures/fig09_ciw5_per_book_test_breakdown.png`
- `outputs/final_report/figures/fig10_contribution_and_use_cases_map.png`
- `outputs/final_report/figures/fig11_feature_cluster_member_trajectories_ma5.png`
- `outputs/final_report/tables/dataset_profile_for_report.csv`
- `outputs/final_report/tables/variant_selection_summary.csv`
- `outputs/final_report/tables/variant_selection_diagnostics.csv`
- `outputs/final_report/tables/ciw5_per_book_deepdive.csv`
- `outputs/final_report/tables/key_results_registry.csv`
- `outputs/final_report/tables/method_claims_checklist.csv`
- `outputs/final_report/tables/cluster_summary_for_report.csv`
- `outputs/final_report/tables/report_integrity_checks.csv`
- `docs/FINAL_REPORT.md`
- `docs/OTHER_EXPERIMENTS.md`

## Known Limitations
- Clustering settings are baseline defaults (`n_clusters=4`) and not heavily tuned.
- DTW hierarchical clustering can produce imbalanced clusters.
- Rank/genre metadata in catalog are useful labels but should not be treated as objective ground truth.
- Insight claims should be validated across parameter sensitivity and method variants.

## Next-Step Roadmap
1. Expand interactive trajectory views (book-level + cluster-level dashboards).
2. Add cluster stability checks across seeds, `k` values, and feature subsets.
3. Build archetype-level summaries with confidence scoring.
4. Extend final-report packaging with sensitivity appendices and versioned claim registries.

# Source: docs/EDA_VIZ_PLAN.md

# EDA and Visualization Plan

## EDA Goals
- Understand global behavior of Twist Signal features across the 20-book corpus.
- Compare cluster structures across methods (`kmeans`, `feature_ward`, `dtw_average`).
- Connect model-derived behavior (`s_t`, `a_t`, peaks, DTW) with catalog metadata labels.
- Produce reusable figures and tables for follow-up research notebooks/reports.

## Analysis Questions
1. Which books show highest novelty and acceleration intensity?
2. How much do feature distributions shift across `k = 5, 7, 11`?
3. Are cluster assignments aligned with metadata labels (genre, format, publication period, origin)?
4. Which books are nearest neighbors by DTW distance?
5. Which books look like outliers under Twist Signal dynamics?
6. How consistent are conclusions across feature-based and DTW-based grouping?

## Visualization Backlog (Priority-Ordered)
1. **Corpus composition**
- Count plots: `genre_primary`, `format`, `origin_country`, `original_language`
- Publication-year histogram

2. **Twist Signal feature distributions**
- Histograms/boxplots/violin by `k` for `mean_s`, `std_s`, `max_s`, `mean_a`, `std_a`, `max_a`
- Correlation heatmap for numeric features
- Interactive scatter(s) for selected feature pairs

3. **Cluster diagnostics**
- Cluster size by method and `k`
- Cluster composition by metadata categories
- Cross-tab agreement: KMeans vs feature_ward

4. **DTW structure**
- DTW heatmap with labels
- Nearest-neighbor table (top 1-3)
- MDS projection from DTW matrix

5. **Book-level trajectory deep dives**
- `s_t` and `a_t` line plots for representative books
- Peak markers overlaid on acceleration
- Metadata and `twist_peak_reason` annotation panel

## Insight Framework
Two interpretation tiers:

1. **Exploratory Observations**
- Descriptive summaries of what is directly visible in current outputs.
- No causal claims.
- Focus on distributions, rankings, and grouping patterns.

2. **Deeper Interpretation (Hypotheses)**
- Candidate narrative archetypes and cross-book storyline behavior patterns.
- Must include confidence level and caveat line for each claim.

Confidence labels:
- `High`: pattern appears across methods/parameters and has strong separation.
- `Medium`: pattern appears in one method or one parameterization with moderate support.
- `Low`: tentative pattern requiring further validation.

## Validation Before Claiming Insights

Required checks before stronger claims:
1. Compare conclusions across `k = 5, 7, 11`.
2. Compare KMeans vs hierarchical outcomes.
3. Re-check key findings with and without selected metadata fields.
4. Track outlier sensitivity (books with very high/low `T` or extreme max values).
5. Validate DTW findings against feature-space findings.

Optional advanced checks:
- Bootstrap sampling on feature rows.
- Alternative distance metrics (cosine/euclidean on normalized time series).
- Stability diagnostics over random seeds for clustering.

## Deliverables and Iteration Loop
Iteration loop:
1. Run `03_eda_and_visualization.ipynb`.
2. Export figures/tables to `outputs/eda/`.
3. Review `outputs/eda/insights.md`.
4. Convert low-confidence interpretations into explicit follow-up tests.

Primary deliverables:
- `outputs/eda/figures/*`
- `outputs/eda/tables/*`
- `outputs/eda/insights.md`

Metadata-guided analyses to always include:
- Categorical: `genre_primary`, `format`, `origin_country`, `original_language`
- Temporal: `first_publication_year`
- Human-labeled ranks: `recognizability_rank`, `genre_clarity_rank`, `twist_peak_rank`
- Relation between these labels and Twist Signal features + cluster assignments

---

# Source: docs/NOVEL_STACKED_OUTPUT_INTERPRETATION.md

# Novel Stacked Twist Signal Interpretation

## Overview
This document interprets 20 stacked per-novel plots generated from `k=5,7,11`. Each figure has three vertical panels for one novel, and each panel overlays `s_t` (Twist Signal) and `a_t` (Twist Acceleration).

How to read each panel:
- `s_t` tracks novelty versus recent narrative context.
- `a_t` tracks local novelty acceleration between consecutive chunks.
- Peak markers indicate top acceleration points for that `k`.

## Global Highlights

Top 3 by novelty (`mean_s`, k=7):
- 35 | The Time Machine | mean_s_k7=0.222
- 345 | Dracula | mean_s_k7=0.222
- 175 | The Phantom of the Opera | mean_s_k7=0.210

Lowest 3 by novelty (`mean_s`, k=7):
- 1342 | Pride and Prejudice | mean_s_k7=0.142
- 1257 | The Three Musketeers | mean_s_k7=0.164
- 768 | Wuthering Heights | mean_s_k7=0.165

Top 3 by acceleration spikes (`max_a`, k=7):

- 84 | Frankenstein; Or, The Modern Prometheus | max_a_k7=0.491
- 113 | The Secret Garden | max_a_k7=0.467
- 43 | The Strange Case of Dr. Jekyll and Mr. Hyde | max_a_k7=0.434

Strongest k-sensitivity (`|delta_mean_s_k11_k5|`):
- 84 | Frankenstein; Or, The Modern Prometheus | delta_mean_s_k11_k5=0.025
- 35 | The Time Machine | delta_mean_s_k11_k5=0.024
- 521 | Robinson Crusoe | delta_mean_s_k11_k5=0.023

Strongest k-sensitivity (`|delta_max_a_k11_k5|`):
- 175 | The Phantom of the Opera | delta_max_a_k11_k5=-0.064
- 103 | Around the World in Eighty Days | delta_max_a_k11_k5=-0.061
- 1513 | Romeo and Juliet | delta_max_a_k11_k5=-0.054

Caveats:
- Labels are relative to this 20-book corpus and current embedding/signal settings.
- Peak extraction uses top-3 acceleration peaks and minimum separation defaults from the pipeline.
- Interpretive statements are descriptive and should be validated with additional settings/checks.

## Per-Novel Interpretations

### [11] Alice's Adventures in Wonderland

![Alice's Adventures in Wonderland stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_11_alice_s_adventures_wonderland_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.202): **High**
- Acceleration/volatility level (k=7 max_a=0.361): **Medium**
- Peak timing profile (k=7): Early-weighted (avg=0.19)
- k-dependence pattern: Increasing with larger k (delta_mean_s=0.019, delta_max_a=0.014)
- What stands out: Alice's Adventures in Wonderland is consistently novel but less spike-driven than the most volatile books.

### [16] Peter Pan

![Peter Pan stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_16_peter_pan_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.202): **Medium**
- Acceleration/volatility level (k=7 max_a=0.233): **Low**
- Peak timing profile (k=7): Early-weighted (avg=0.14)
- k-dependence pattern: Increasing with larger k (delta_mean_s=0.019, delta_max_a=0.017)
- What stands out: Peter Pan sits in a middle regime with increasing with larger k behavior across window sizes.

### [35] The Time Machine

![The Time Machine stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_35_time_machine_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.222): **High**
- Acceleration/volatility level (k=7 max_a=0.414): **High**
- Peak timing profile (k=7): Early-weighted (avg=0.33)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.024, delta_max_a=0.000)
- What stands out: The Time Machine combines high novelty and sharp shifts, forming a jagged trajectory profile.

### [36] The War of the Worlds

![The War of the Worlds stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_36_war_worlds_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.202): **Medium**
- Acceleration/volatility level (k=7 max_a=0.277): **Low**
- Peak timing profile (k=7): Mid-story weighted (avg=0.60)
- k-dependence pattern: Increasing with larger k (delta_mean_s=0.017, delta_max_a=0.040)
- What stands out: The War of the Worlds sits in a middle regime with increasing with larger k behavior across window sizes.

### [43] The Strange Case of Dr. Jekyll and Mr. Hyde

![The Strange Case of Dr. Jekyll and Mr. Hyde stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_43_strange_case_dr_jekyll_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.177): **Low**
- Acceleration/volatility level (k=7 max_a=0.434): **High**
- Peak timing profile (k=7): Distributed across early-to-late arc (avg=0.32)
- k-dependence pattern: Stable across k (delta_mean_s=0.010, delta_max_a=0.000)
- What stands out: The Strange Case of Dr. Jekyll and Mr. Hyde has a calmer baseline punctuated by concentrated bursts of change.

### [55] The Wonderful Wizard of Oz

![The Wonderful Wizard of Oz stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_55_wonderful_wizard_oz_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.175): **Low**
- Acceleration/volatility level (k=7 max_a=0.296): **Low**
- Peak timing profile (k=7): Distributed across early-to-late arc (avg=0.41)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.021, delta_max_a=0.000)
- What stands out: The Wonderful Wizard of Oz shows a smoother, lower-volatility progression relative to the corpus.

### [84] Frankenstein; Or, The Modern Prometheus

![Frankenstein; Or, The Modern Prometheus stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_84_frankenstein_modern_prometheus_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.204): **High**
- Acceleration/volatility level (k=7 max_a=0.491): **High**
- Peak timing profile (k=7): Early-weighted (avg=0.27)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.025, delta_max_a=0.000)
- What stands out: Frankenstein; Or, The Modern Prometheus combines high novelty and sharp shifts, forming a jagged trajectory profile.

### [103] Around the World in Eighty Days

![Around the World in Eighty Days stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_103_around_world_eighty_days_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.191): **Medium**
- Acceleration/volatility level (k=7 max_a=0.259): **Low**
- Peak timing profile (k=7): Distributed across early-to-late arc (avg=0.48)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.023, delta_ma

x_a=-0.061)
- What stands out: Around the World in Eighty Days sits in a middle regime with mixed sensitivity across k behavior across window sizes.

### [113] The Secret Garden

![The Secret Garden stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_113_secret_garden_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.194): **Medium**
- Acceleration/volatility level (k=7 max_a=0.467): **High**
- Peak timing profile (k=7): Distributed across early-to-late arc (avg=0.45)
- k-dependence pattern: Stable across k (delta_mean_s=0.015, delta_max_a=0.000)
- What stands out: The Secret Garden sits in a middle regime with stable across k behavior across window sizes.

### [120] Treasure Island

![Treasure Island stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_120_treasure_island_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.188): **Medium**
- Acceleration/volatility level (k=7 max_a=0.327): **Medium**
- Peak timing profile (k=7): Early-weighted (avg=0.25)
- k-dependence pattern: Stable across k (delta_mean_s=0.012, delta_max_a=0.000)
- What stands out: Treasure Island sits in a middle regime with stable across k behavior across window sizes.

### [175] The Phantom of the Opera

![The Phantom of the Opera stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_175_phantom_opera_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.210): **High**
- Acceleration/volatility level (k=7 max_a=0.363): **High**
- Peak timing profile (k=7): Distributed across early-to-late arc (avg=0.51)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.022, delta_max_a=-0.064)
- What stands out: The Phantom of the Opera combines high novelty and sharp shifts, forming a jagged trajectory profile.

### [345] Dracula

![Dracula stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_345_dracula_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.222): **High**
- Acceleration/volatility level (k=7 max_a=0.383): **High**
- Peak timing profile (k=7): Early-weighted (avg=0.05)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.021, delta_max_a=-0.003)
- What stands out: Dracula combines high novelty and sharp shifts, forming a jagged trajectory profile.

### [521] Robinson Crusoe

![Robinson Crusoe stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_521_robinson_crusoe_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.207): **High**
- Acceleration/volatility level (k=7 max_a=0.386): **High**
- Peak timing profile (k=7): Early-weighted (avg=0.19)

- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.023, delta_max_a=0.000)
- What stands out: Robinson Crusoe combines high novelty and sharp shifts, forming a jagged trajectory profile.

### [768] Wuthering Heights

![Wuthering Heights stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_768_wuthering_heights_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.165): **Low**
- Acceleration/volatility level (k=7 max_a=0.317): **Medium**
- Peak timing profile (k=7): Late-weighted (avg=0.81)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.006, delta_max_a=-0.047)
- What stands out: Wuthering Heights sits in a middle regime with mixed sensitivity across k behavior across window sizes.

### [1184] The Count of Monte Cristo

![The Count of Monte Cristo stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_1184_count_monte_cristo_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.176): **Low**
- Acceleration/volatility level (k=7 max_a=0.307): **Low**
- Peak timing profile (k=7): Distributed across early-to-late arc (avg=0.50)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.017, delta_max_a=-0.006)
- What stands out: The Count of Monte Cristo shows a smoother, lower-volatility progression relative to the corpus.

### [1257] The Three Musketeers

![The Three Musketeers stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_1257_three_musketeers_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.164): **Low**
- Acceleration/volatility level (k=7 max_a=0.320): **Medium**
- Peak timing profile (k=7): Early-weighted (avg=0.23)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.014, delta_max_a=0.031)
- What stands out: The Three Musketeers sits in a middle regime with mixed sensitivity across k behavior across window sizes.

### [1260] Jane Eyre

![Jane Eyre stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_1260_jane_eyre_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.196): **Medium**
- Acceleration/volatility level (k=7 max_a=0.344): **Medium**
- Peak timing profile (k=7): Distributed across early-to-late arc (avg=0.45)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.012, delta_max_a=0.042)
- What stands out: Jane Eyre sits in a middle regime with mixed sensitivity across k behavior across window sizes.

### [1342] Pride and Prejudice

![Pride and Prejudice stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_1342_pride_prejudice_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.142): **Low**
- Acceleration/volatility level (k=7 max_a=0.229): **Low**
- Peak timing profile (k=7): Mid-story weighted (avg=0.62)
- k-dependence pattern: Stable across k (delta_mean_s=0.009, delta_max_a=-0.001)
- What stands out: Pride and Prejudice shows a smoother, lower-volatility progression relative to the corpus.

### [1513] Romeo and Juliet

![Romeo and Juliet stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_1513_romeo_juliet_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.175): **Low**
- Acceleration/volatility level (k=7 max_a=0.208): **Low**
- Peak timing profile (k=7): Distributed across early-to-late arc (avg=0.46)
- k-dependence pattern: Mixed sensitivity across k (delta_mean_s=0.015, delta_max_a=-0.054)
- What stands out: Romeo and Juliet shows a smoother, lower-volatility progression relative to the corpus.

### [1661] The Adventures of Sherlock Holmes

![The Adventures of Sherlock Holmes stacked Twist Signal](../outputs/eda/novel_stacks/figures/novel_1661_adventures_sherlock_holmes_stacked_k5_k7_k11.png)

- Novelty level (k=7 mean_s=0.208): **High**
- Acceleration/volatility level (k=7 max_a=0.329): **Medium**
- Peak timing profile (k=7): Mid-story weighted (avg=0.40)
- k-dependence pattern: Increasing with larger k (delta_mean_s=0.016, delta_max_a=0.016)
- What stands out: The Adventures of Sherlock Holmes is consistently novel but less spike-driven than the most volatile books.

# Source: docs/OUTPUT_INTERPRETATION.md

# Output Interpretation Guide

## Scope
This document interprets all generated EDA plot outputs in `outputs/eda/figures/`, using supporting tables in `outputs/eda/tables/` and current run artifacts.

Run context:
- Books: 20
- Feature rows: 60 (`k in [5, 7, 11]`)
- Twist Signal model in current processed indexes: `sentence-transformers/all-mpnet-base-v2`

## High-Level Reading
Main takeaways from the current outputs:
1. The corpus is diverse but UK-heavy and novel-heavy.
2. Twist Signal feature intensity tends to increase from `k=5` to `k=11` for novelty (`mean_s`, `max_s`), while acceleration (`mean_a`) remains close.
3. Feature-based clustering is moderately structured and reasonably balanced; DTW hierarchical clustering is highly imbalanced (`17/1/1/1`).
4. Deep-dive representatives show different trajectory regimes: long gradual arcs vs short high-volatility trajectories.

## Plot-by-Plot Interpretation

### Corpus Composition
1. `corpus_composition_counts.png`
- What it shows: counts by `genre_primary`, `format`, `origin_country`, `original_language`.
- Interpretation:
  - `Adventure` is the largest genre block (5 books), then `Fantasy/Sci-Fi/Gothic` (3 each).
  - Format is dominated by `novel` (16/20), with sparse `novella/play/short-story collection`.
  - Country composition is UK-heavy (13/20), then France (4), US (2), Ireland (1).
  - Language split is English (16) vs French (4).
- Implication: downstream cluster differences can reflect corpus composition bias, not just narrative dynamics.
![Corpus composition counts](../outputs/eda/figures/corpus_composition_counts.png)

2. `publication_year_distribution.png`
- What it shows: publication-year histogram.
- Interpretation: wide temporal range from 1597 to 1911 (median ~1877.5), concentrated in the 19th century.
- Implication: year effects may confound stylistic/trajectory effects; include year in stratified analyses.
![Publication year distribution](../outputs/eda/figures/publication_year_distribution.png)

3. `metadata_rank_distributions.png`
- What it shows: distributions of `recognizability_rank`, `genre_clarity_rank`, `twist_peak_rank`.
- Interpretation: all three ranks are evenly spread 1-20 by design (mean/median ~10.5).
- Implication: these are balanced labels and useful for comparative slicing, but still human-assigned.
![Metadata rank distributions](../outputs/eda/figures/metadata_rank_distributions.png)

### Twist Signal Feature EDA
4. `twist_signal_feature_hist_by_k.png`
- What it shows: per-feature distributions across `k=5,7,11`.
- Interpretation:
  - `mean_s` shifts upward with larger `k` (means: `0.1829 -> 0.1910 -> 0.1998`).
  - `max_s` also increases with `k` on average.
  - `mean_a` remains in a narrow band (~0.047-0.048).
- Implication: context window changes baseline novelty scale; compare books within the same `k`.
![Twist Signal feature histograms by k](../outputs/eda/figures/twist_signal_feature_hist_by_k.png)

5. `twist_signal_feature_boxplots_by_k.png`
- What it shows: spread and outliers by `k`.
- Interpretation:
  - `max_a` has broad spread and strongest outlier behavior.
  - `std_s` and `std_a` retain meaningful cross-book separation across all `k`.
- Implication: volatility-oriented features are likely key drivers for clustering.
![Twist Signal feature boxplots by k](../outputs/eda/figures/twist_signal_feature_boxplots_by_k.png)

6. `twist_signal_feature_correlation_heatmap.png`
- What it shows: correlation structure across numeric Twist Signal features.

- Interpretation (strongest absolute relationships):
  - `mean_a` vs `std_a`: `0.889`
  - `std_s` vs `mean_a`: `0.819`
  - `mean_s` vs `std_s`: `0.719`
  - `mean_s` vs `max_s`: `0.670`
- Implication: novelty and acceleration are linked but non-redundant; keep both families in feature space.
![Twist Signal feature correlation heatmap](../outputs/eda/figures/twist_signal_feature_correlation_heatmap.png)

7. `interactive_twist_signal_scatter_k7.html`
- What it shows: `mean_s` vs `std_a` at `k=7`, colored by genre and symbolized by format.
- Interpretation:
  - Positive trend (`corr ~ 0.567`) indicates books with higher average novelty often have higher acceleration variability.
  - Upper-right region includes `The Time Machine` and `Dracula`, indicating high novelty and volatility.
- Implication: this view is a strong first-pass outlier/archetype selector for deep dives.
- Interactive view: [interactive_twist_signal_scatter_k7.html](../outputs/eda/figures/interactive_twist_signal_scatter_k7.html)

### Cluster Analysis
8. `cluster_sizes_kmeans_vs_hier_feature_ward.png`
- What it shows: cluster counts by `k` for KMeans and hierarchical feature-ward.
- Interpretation:
  - KMeans stays fairly balanced across `k`.
  - Feature-ward is also workable but shows mild imbalance at some `k` values.
- Implication: feature-space clustering is usable for archetype exploration.
![Cluster sizes: KMeans vs hierarchical feature-ward](../outputs/eda/figures/cluster_sizes_kmeans_vs_hier_feature_ward.png)

9. `kmeans_k7_cluster_genre_composition.png`
- What it shows: genre mix per KMeans cluster at `k=7`.
- Interpretation:
  - Cluster 3 is Fantasy-heavy.
  - Cluster 2 mixes Horror/Sci-Fi with some Adventure/Gothic/Mystery.
  - No cluster is pure single-genre overall.
- Implication: Twist Signal captures pacing/trajectory structure that crosses genre boundaries.
![KMeans k7 cluster genre composition](../outputs/eda/figures/kmeans_k7_cluster_genre_composition.png)

10. `agreement_heatmap_k5_kmeans_vs_ward.png`
11. `agreement_heatmap_k7_kmeans_vs_ward.png`
12. `agreement_heatmap_k11_kmeans_vs_ward.png`
- What they show: cross-tab agreement between KMeans and feature-ward per `k`.
- Interpretation:
  - Agreement is moderate, not trivial:
    - `k=5`: ARI `0.481`, NMI `0.650`
    - `k=7`: ARI `0.398`, NMI `0.628`
    - `k=11`: ARI `0.470`, NMI `0.624`
- Implication: there is consistent shared structure, but cluster boundaries are method-sensitive.
![Agreement heatmap k=5](../outputs/eda/figures/agreement_heatmap_k5_kmeans_vs_ward.png)
![Agreement heatmap k=7](../outputs/eda/figures/agreement_heatmap_k7_kmeans_vs_ward.png)
![Agreement heatmap k=11](../outputs/eda/figures/agreement_heatmap_k11_kmeans_vs_ward.png)

```
### DTW Structure
13. `dtw_heatmap_k7.html`
- What it shows: pairwise DTW distances on resampled Twist Signal (`k=7`).
- Interpretation:
  - Global mean distance `~0.618`.
  - Nearest-neighbor mean distance `~0.550` (min `~0.464`).
  - Closest pairs include:
    - Alice ↔ Treasure Island
    - Alice ↔ Romeo and Juliet
    - Treasure Island ↔ Romeo and Juliet
- Implication: some cross-genre books share similar temporal novelty shape.
- Interactive view: [dtw_heatmap_k7.html](../outputs/eda/figures/dtw_heatmap_k7.html)

14. `dtw_mds_projection.png`
15. `interactive_dtw_mds_projection.html`
- What they show: 2D MDS embedding from DTW distances.
- Interpretation:
  - A dominant central basin appears, consistent with DTW cluster collapse.
  - Limited separation for most books under current DTW + average linkage settings.
- Implication: DTW representation likely needs alternative linkage/normalization or richer sequence features for balanced clustering.
![DTW MDS projection](../outputs/eda/figures/dtw_mds_projection.png)
- Interactive view: [interactive_dtw_mds_projection.html](../outputs/eda/figures/interactive_dtw_mds_projection.html)

### Book-Level Deep Dives (k=7 Representatives)
16. `twist_signal_deep_dive_book_1184_k7.png` (Cluster 0, *The Count of Monte Cristo*)
- Interpretation:
  - Very long trajectory (`T=4608`), moderate novelty (`mean_s=0.176`) and smoother acceleration (`max_a=0.307`).
  - Peaks spread across narrative (indices ~786, 2681, 3391), consistent with long-arc progression.
![Deep dive: 1184](../outputs/eda/figures/twist_signal_deep_dive_book_1184_k7.png)

17. `twist_signal_deep_dive_book_36_k7.png` (Cluster 1, *The War of the Worlds*)
- Interpretation:
  - Medium length (`T=599`), high novelty (`mean_s=0.202`) and moderate-high acceleration.
  - Peaks in mid-late sections (~278, 371, 431), showing concentrated escalation zones.
![Deep dive: 36](../outputs/eda/figures/twist_signal_deep_dive_book_36_k7.png)

18. `twist_signal_deep_dive_book_35_k7.png` (Cluster 2, *The Time Machine*)
- Interpretation:
  - Shorter trajectory (`T=323`) with very high novelty and volatility (`mean_s=0.222`, `max_a=0.415`).
  - Early first peak plus later major peaks indicates rapid regime shifts.
![Deep dive: 35](../outputs/eda/figures/twist_signal_deep_dive_book_35_k7.png)

19. `twist_signal_deep_dive_book_11_k7.png` (Cluster 3, *Alice's Adventures in Wonderland*)
- Interpretation:
  - Short trajectory (`T=264`), high novelty (`mean_s=0.202`) but lower average acceleration than cluster-2 representative.
  - Early/mid peaks suggest episodic local shifts rather than sustained escalation.
![Deep dive: 11](../outputs/eda/figures/twist_signal_deep_dive_book_11_k7.png)
```

## Cross-Output Interpretation Notes

1. Feature-space methods surface multiple trajectory archetypes; DTW-average currently under-separates the corpus.
2. High `mean_s` does not always imply highest `max_a`; novelty level and volatility should be interpreted separately.
3. Metadata labels (genre/ranks) are useful context but not strict validators of trajectory clusters.

## Confidence and Caveats

- Confidence: **Medium** for descriptive pattern claims (distribution, cluster size, correlation, nearest-neighbor structure).
- Confidence: **Low-to-Medium** for archetype-level narrative interpretations without external validation.

Known caveats:

1. Default clustering hyperparameters were not optimized.
2. DTW was computed on resampled univariate Twist Signal only.
3. Corpus composition imbalance (genre/country/format) can affect apparent structure.

## Recommended Next Validation Pass

1. Compare cluster quality metrics for alternative `n_clusters` and linkage choices.
2. Add DTW variants (derivative DTW, z-normalized per-book signal, multivariate distance with `s_t` and `a_t`).
3. Evaluate stability across seeds and feature subsets.
4. Add interpretable temporal motifs (e.g., early/mid/late segment stats) to complement global aggregates.

# Source: outputs/eda/insights.md

# EDA Insights: Twist Signal

## Exploratory Observations

- k=7 mean_s ranges from 0.142 to 0.222 across books.
- k=7 max_a ranges from 0.208 to 0.491.
- DTW distance matrix mean is 0.618 with nearest-neighbor mean 0.550.
- Feature-based clusters are more balanced than DTW-average hierarchical clusters in current settings.

Top books by mean_s (k=7):

- 35 | The Time Machine | mean_s=0.222 | genre=Sci-Fi
- 345 | Dracula | mean_s=0.222 | genre=Horror
- 175 | The Phantom of the Opera | mean_s=0.210 | genre=Gothic

Top books by max_a (k=7):

- 84 | Frankenstein; Or, The Modern Prometheus | max_a=0.491 | genre=Sci-Fi
- 113 | The Secret Garden | max_a=0.467 | genre=Children's Fiction
- 43 | The Strange Case of Dr. Jekyll and Mr. Hyde | max_a=0.434 | genre=Horror

Low mean_s books (k=7):

- 1342 | Pride and Prejudice | mean_s=0.142 | genre=Romance
- 1257 | The Three Musketeers | mean_s=0.164 | genre=Adventure
- 768 | Wuthering Heights | mean_s=0.165 | genre=Gothic

## Deeper Interpretation Hypotheses

- Hypothesis (Medium confidence): books with higher max_a show sharper local narrative transitions and may correspond to stronger labeled twist intensity.
- Hypothesis (Medium confidence): feature-space clusters separate multiple traje

ctory regimes, while DTW-average currently collapses many books into a dominant basin.
- Hypothesis (Low confidence): genre labels partially align with trajectory clusters, but overlap suggests shared pacing motifs across genres.

## Caveats and Validation Next Steps
- Caveat: clustering defaults are baseline and not hyperparameter-optimized.
- Caveat: interpretations should be rechecked across k values and alternative distance/linkage settings.
- Caveat: metadata rank labels are helpful but subjective and should not be treated as strict ground truth.

# Source: outputs/eda/novel_stacks/peak_chunk_previews_k5.md

# Peak Chunk Previews (k=5)

- Rows: 60
- Books covered: 20

## [11] Alice's Adventures in Wonderland

- Peak 1: index=1, pos_norm=0.0038, words=100-400
  - Preview: sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, “and what is the use of a book,” thought Alice “without p
- Peak 2: index=48, pos_norm=0.1825, words=4800-5100
  - Preview: my dear?” it continued, turning to Alice as it spoke. “As wet as ever,” said Alice in a melancholy tone: “it doesn’t seem to dry me at all.” “In that case,” said the Dodo solemnly, rising to its feet, “I move that the me
- Peak 3: index=93, pos_norm=0.3536, words=9300-9600
  - Preview: to sell you a couple?” “You are old,” said the youth, “and your jaws are too weak For anything tougher than suet; Yet you finished the goose, with the bones and the beak– Pray, how did you manage to do it?” “In my youth,

## [16] Peter Pan

- Peak 1: index=1, pos_norm=0.0021, words=100-400
  - Preview: Chapter VIII. THE MERMAIDS’ LAGOON Chapter IX. THE NEVER BIRD Chapter X. THE HAPPY HOME Chapter XI. WENDY’S STORY Chapter XII. THE CHILDREN ARE CARRIED OFF Chapter XIII. DO YOU BELIEVE IN FAIRIES? Chapter XIV. THE PIRATE
- Peak 2: index=15, pos_norm=0.0319, words=1500-1800
  - Preview: asleep to rummage in their minds and put things straight for next morning, repacking into their proper places the many articles that have wandered during the day. If you could keep awake (but of course you can’t) you wou
- Peak 3: index=185, pos_norm=0.3936, words=18500-18800
  - Preview: cried Peter sternly. Quickly they made-believe to grow the loveliest roses up the walls. Babies? To prevent Peter ordering babies they hurried into song again: “We’ve made the roses peeping out, The babes are at the door

## [35] The Time Machine

- Peak 1: index=1, pos_norm=0.0031, words=100-400
  - Preview: pale grey eyes shone and twinkled, and his usually pale face was flushed and animated. The fire burnt brightly, and the soft radiance of the incandescent lights in the lilies of silver caught the bubbles that flashed and
- Peak 2: index=113, pos_norm=0.3509, words=11300-11600
  - Preview: down my cheek and chin. All the time I ran I was saying to myself: ‘They have moved it a little, pushed it under the bushes out of the way.’ Nevertheless, I ran with all my might. All the time, with the certainty that so

```
- Peak 3: index=205, pos_norm=0.6366, words=20500-20800
  - Preview: in sight of the palace, silhouetted black against the pale yellow o
f the sky. “Weena had been hugely delighted when I began to carry her, but after
 a while she desired me to let her down, and ran along by the side of me,

## [36] The War of the Worlds

- Peak 1: index=123, pos_norm=0.2057, words=12300-12600
  - Preview: knew, that his wife had gone to London with him and had locked up t
heir house. I went in again, according to my promise, to get my servant’s box, l
ugged it out, clapped it beside her on the tail of the dog cart, and then
- Peak 2: index=371, pos_norm=0.6204, words=37100-37400
  - Preview: how this change affected our position, save that we were relieved o
f our fear of the Black Smoke. But later I perceived that we were no longer hemm
ed in, that now we might get away. So soon as I realised that the way of
- Peak 3: index=444, pos_norm=0.7425, words=44400-44700
  - Preview: the sky to herself. I heard a dog howling, and that familiar sound
it was that made me listen. Then I heard quite distinctly a booming exactly like
 the sound of great guns. Six distinct reports I counted, and after a lon

## [43] The Strange Case of Dr. Jekyll and Mr. Hyde

- Peak 1: index=1, pos_norm=0.0039, words=100-400
  - Preview: when the wine was to his taste, something eminently human beaconed
from his eye; something indeed which never found its way into his talk, but whic
h spoke not only in these silent symbols of the after-dinner face, but mo
- Peak 2: index=60, pos_norm=0.2362, words=6000-6300
  - Preview: the fire. “I have no doubt you are perfectly right,” he said at las
t, getting to his feet. “Well, but since we have touched upon this business, and
 for the last time I hope,” continued the doctor, “there is one point I s
- Peak 3: index=184, pos_norm=0.7244, words=18400-18700
  - Preview: at the table and held on, staring with injected eyes, gasping with
open mouth; and as I looked there came, I thought, a change—he seemed to swell—h
is face became suddenly black and the features seemed to melt and alter—a

## [55] The Wonderful Wizard of Oz

- Peak 1: index=4, pos_norm=0.0102, words=400-700
  - Preview: Em, who was the farmer’s wife. Their house was small, for the lumbe
r to build it had to be carried by wagon many miles. There were four walls, a fl
oor and a roof, which made one room; and this room contained a rusty look
- Peak 2: index=122, pos_norm=0.3096, words=12200-12500
  - Preview: a tree full of fine fruit. This pleased Dorothy, who had eaten noth
ing but nuts all day, and she made a hearty meal of the ripe fruit. But it takes
 time to make a raft, even when one is as industrious and untiring as the
- Peak 3: index=362, pos_norm=0.9188, words=36200-36500
  - Preview: I could carry you in my basket.” “That would make me very unhappy,”
 answered the china Princess. “You see, here in our country we live contentedly,
 and can talk and move around as we please. But whenever any of us are ta

## [84] Frankenstein; Or, The Modern Prometheus

- Peak 1: index=1, pos_norm=0.0013, words=100-400
  - Preview: forebodings. I arrived here yesterday, and my first task is to assu
re my dear sister of my welfare and increasing confidence in the success of my u
ndertaking. I am already far north of London, and as I walk in the street
- Peak 2: index=146, pos_norm=0.1952, words=14600-14900
  - Preview: before. Winter, spring, and summer passed away during my labours; b
ut I did not watch the blossom or the expanding leaves—sights which before alway
s yielded me supreme delight—so deeply was I engrossed in my occupation.
- Peak 3: index=307, pos_norm=0.4104, words=30700-31000
  - Preview: origin and author? Cursed be the day, abhorred devil, in which you
```

first saw light! Cursed (although I curse myself) be the hands that formed you! You have made me wretched beyond expression. You have left me no power to

## [103] Around the World in Eighty Days

- Peak 1: index=428, pos_norm=0.6783, words=42800-43100
  - Preview: station drew up to the door. As he was getting in, Mr. Fogg said to Fix, “You have not seen this Colonel Proctor again?” “No.” “I will come back to America to find him,” said Phileas Fogg calmly. “It would not be right f
- Peak 2: index=434, pos_norm=0.6878, words=43400-43700
  - Preview: at New York on the 11th for Liverpool. The car which he occupied was a sort of long omnibus on eight wheels, and with no compartments in the interior. It was supplied with two rows of seats, perpendicular to the directio
- Peak 3: index=582, pos_norm=0.9223, words=58200-58500
  - Preview: have dry wood to keep the steam up to the adequate pressure, and on that day the poop, cabins, bunks, and the spare deck were sacrificed. On the next day, the 19th of December, the masts, rafts, and spars were burned; th

## [113] The Secret Garden

- Peak 1: index=1, pos_norm=0.0012, words=100-400
  - Preview: A BIT OF EARTH?” XIII. “I AM COLIN” XIV. A YOUNG RAJAH XV. NEST BUILDING XVI. “I WON’T!” SAID MARY XVII. A TANTRUM XVIII. “THA’ MUNNOT WASTE NO TIME” XIX. “IT HAS COME!” XX. “I SHALL LIVE FOREVER—AND EVER—AND EVER!” XXI.
- Peak 2: index=374, pos_norm=0.4652, words=37400-37700
  - Preview: Magic in India, but I can’t make it. I just went into his room and I was so surprised to see him I stood and stared. And then he turned round and stared at me. And he thought I was a ghost or a dream and I thought perhap
- Peak 3: index=711, pos_norm=0.8843, words=71100-71400
  - Preview: content. Fears for the Eggs became things of the past. Knowing that your Eggs were as safe as if they were locked in a bank vault and the fact that you could watch so many curious things going on made setting a most ente

## [120] Treasure Island

- Peak 1: index=1, pos_norm=0.0015, words=100-400
  - Preview: fall on! If not, If studious youth no longer crave, His ancient appetites forgot, Kingston, or Ballantyne the brave, Or Cooper of the wood and wave: So be it, also! And may I And all my pirates share the grave Where thes
- Peak 2: index=59, pos_norm=0.0863, words=5900-6200
  - Preview: if I can’t get away nohow, and they tip me the black spot, mind you, it’s my old sea-chest they’re after; you get on a horse--you can, can’t you? Well, then, you get on a horse, and go to--well, yes, I will!--to that ete
- Peak 3: index=444, pos_norm=0.6491, words=44400-44700
  - Preview: the narrows for the open sea. Suddenly the schooner in front of me gave a violent yaw, turning, perhaps, through twenty degrees; and almost at the same moment one shout followed another from on board; I could hear feet p

## [175] The Phantom of the Opera

- Peak 1: index=49, pos_norm=0.0574, words=4900-5200
  - Preview: is quite authentic, from M. Pedro Gailhard himself, the late manager of the Opera. Chapter II The New Margarita On the first landing, Sorelli ran against the Comte de Chagny, who was coming up-stairs. The count, who was
- Peak 2: index=55, pos_norm=0.0645, words=5500-5800
  - Preview: had often said she meant to practise alone for the future. The whole thing was a mystery. The Comte de Chagny, standing up in his box, listened to all this frenzy and took part in it by loudly applauding. Philippe George
- Peak 3: index=663, pos_norm=0.7773, words=66300-66600
  - Preview: we were going to fight a duel. I said: "Yes; and what a duel!" But, of course, I had no time to explain anything to him. The little viscount is a brave fellow, but he knew hardly anything about his adversary; and it was

```
## [345] Dracula

- Peak 1: index=128, pos_norm=0.0795, words=12800-13100
  - Preview: days are over. Blood is too precious a thing in these days of disho
nourable peace; and the glories of the great races are as a tale that is told.”
It was by this time close on morning, and we went to bed. (_Mem._, this d
- Peak 2: index=135, pos_norm=0.0838, words=13500-13800
  - Preview: himself on these points of which he had spoken, and I had verified
all as well as I could by the books available, he suddenly stood up and said:--
“Have you written since your first letter to our friend Mr. Peter Hawkins
- Peak 3: index=752, pos_norm=0.4668, words=75200-75500
  - Preview: It was terribly weak, and looked quite emaciated. It too, when part
ially restored, had the common story to tell of being lured away by the “bloofer
 lady.” CHAPTER XIV MINA HARKER’S JOURNAL _23 September_.--Jonathan is be

## [521] Robinson Crusoe

- Peak 1: index=1, pos_norm=0.0008, words=100-400
  - Preview: CANNIBALS CHAPTER XVII—VISIT OF MUTINEERS CHAPTER XVIII—THE SHIP RE
COVERED CHAPTER XIX—RETURN TO ENGLAND CHAPTER XX—FIGHT BETWEEN FRIDAY AND A BEAR
 CHAPTER I. START IN LIFE I was born in the year 1632, in the city of Yor
- Peak 2: index=378, pos_norm=0.3132, words=37800-38100
  - Preview: bestirred myself to furnish myself with everything that I wanted, a
nd make my way of living as regular as I could. From the 4th of July to the 14th
 I was chiefly employed in walking about with my gun in my hand, a little
- Peak 3: index=443, pos_norm=0.3670, words=44300-44600
  - Preview: in England, without any order of mine, to pack it up among my goods
, and for assisting me afterwards to save it out of the wreck of the ship. Thus,
 and in this disposition of mind, I began my third year; and though I hav

## [768] Wuthering Heights

- Peak 1: index=727, pos_norm=0.6283, words=72700-73000
  - Preview: village? She said he had only been twice, on horseback, accompanyin
g his father; and both times he pretended to be quite knocked up for three or fo
ur days afterwards. That housekeeper left, if I recollect rightly, two ye
- Peak 2: index=960, pos_norm=0.8297, words=96000-96300
  - Preview: pony Minny, if I would get the key of our room, and let her out; bu
t I told her she had nothing to give, they were all, all mine. And then she crie
d, and took a little picture from her neck, and said I should have that;
- Peak 3: index=1135, pos_norm=0.9810, words=113500-113800
  - Preview: within two yards’ distance. And whatever it was, it communicated, a
pparently, both pleasure and pain in exquisite extremes: at least the anguished,
 yet raptured, expression of his countenance suggested that idea. The fan

## [1184] The Count of Monte Cristo

- Peak 1: index=786, pos_norm=0.1706, words=78600-78900
  - Preview: wall in characters of flame—if he slept for a moment the wildest dr
eams haunted his brain. He ascended into grottos paved with emeralds, with panel
s of rubies, and the roof glowing with diamond stalactites. Pearls fell d
- Peak 2: index=2681, pos_norm=0.5819, words=268100-268400
  - Preview: rainbow when it disappears, pass through all the prismatic shades,
after which they were sent to the kitchen. Their agony formed part of their meri
t—if they were not seen alive, they were despised when dead.” “Yes,” said
- Peak 3: index=3391, pos_norm=0.7361, words=339100-339400
  - Preview: meteor seen passing inauspiciously between two clouds in a stormy s
ky. Chapter 81. The Room of the Retired Baker The evening of the day on which th
e Count of Morcerf had left Danglars’ house with feelings of shame and an

## [1257] The Three Musketeers
```

```
- Peak 1: index=1, pos_norm=0.0004, words=100-400
  - Preview: THICKENS Chapter XII. GEORGE VILLIERS, DUKE OF BUCKINGHAM Chapter X
III. MONSIEUR BONACIEUX Chapter XIV. THE MAN OF MEUNG Chapter XV. MEN OF THE ROB
E AND MEN OF THE SWORD Chapter XVI. IN WHICH M. SÉGUIER, KEEPER OF THE SE
- Peak 2: index=601, pos_norm=0.2622, words=60100-60400
  - Preview: tempting demon, to have recourse to the bell rope, and ring with al
l his might. At the denunciating sound, the monks would be rendered aware that t
emptation was besieging a brother, and all the community would go to pray
- Peak 3: index=994, pos_norm=0.4337, words=99400-99700
  - Preview: that is a serious infraction of the rule—for eggs are meat, since t
hey engender chickens.” “This feast is not very succulent; but never mind, I wil
l put up with it for the sake of remaining with you.” “I am grateful to y

## [1260] Jane Eyre

- Peak 1: index=1, pos_norm=0.0005, words=100-400
  - Preview: Press, for the fair field its honest suffrage has opened to an obsc
ure aspirant. To my Publishers, for the aid their tact, their energy, their prac
tical sense and frank liberality have afforded an unknown and unrecommend
- Peak 2: index=1245, pos_norm=0.6726, words=124500-124800
  - Preview: it by any legal proceedings: for the doctors now discovered that _m
y wife_ was mad—her excesses had prematurely developed the germs of insanity. Ja
ne, you don’t like my narrative; you look almost sick—shall I defer the r
- Peak 3: index=1252, pos_norm=0.6764, words=125200-125500
  - Preview: was past in a second. “A wind fresh from Europe blew over the ocean
 and rushed through the open casement: the storm broke, streamed, thundered, bla
zed, and the air grew pure. I then framed and fixed a resolution. While I

## [1342] Pride and Prejudice

- Peak 1: index=51, pos_norm=0.0401, words=5100-5400
  - Preview: [Illustration: “He came down to see the place” [_Copyright 1894 by
George Allen._]] This was invitation enough. “Why, my dear, you must know, Mrs.
Long says that Netherfield is taken by a young man of large fortune from
- Peak 2: index=499, pos_norm=0.3926, words=49900-50200
  - Preview: and good conduct, I am sure. You must not disappoint your father.”
“My dear aunt, this is being serious indeed.” “Yes, and I hope to engage you to
be serious likewise.” “Well, then, you need not be under any alarm. I wil
- Peak 3: index=871, pos_norm=0.6853, words=87100-87400
  - Preview: it seemed long, was not long enough to determine her feelings towar
ds _one_ in that mansion; and she lay awake two whole hours, endeavouring to mak
e them out. She certainly did not hate him. No; hatred had vanished long

## [1513] Romeo and Juliet

- Peak 1: index=42, pos_norm=0.1634, words=4200-4500
  - Preview: happy nights to happy days. [_Exeunt._] SCENE IV. A Street. Enter R
omeo, Mercutio, Benvolio, with five or six Maskers; Torch-bearers and others. RO
MEO. What, shall this speech be spoke for our excuse? Or shall we on with
- Peak 2: index=48, pos_norm=0.1868, words=4800-5100
  - Preview: mind the fairies’ coachmakers. And in this state she gallops night
by night Through lovers’ brains, and then they dream of love; O’er courtiers’ kn
ees, that dream on curtsies straight; O’er lawyers’ fingers, who straight
- Peak 3: index=87, pos_norm=0.3385, words=8700-9000
  - Preview: the earth doth live But to the earth some special good doth give; N
or aught so good but, strain’d from that fair use, Revolts from true birth, stum
bling on abuse. Virtue itself turns vice being misapplied, And vice somet

## [1661] The Adventures of Sherlock Holmes

- Peak 1: index=417, pos_norm=0.3998, words=41700-42000
```

  - Preview: laudanum in an attempt to produce the same effects. He found, as so many more have done, that the practice is easier to attain than to get rid of, and for many years he continued to be a slave to the drug, an object of m
- Peak 2: index=540, pos_norm=0.5177, words=54000-54300
  - Preview: shade instead of ruby red. In spite of its youth, it has already a sinister history. There have been two murders, a vitriol-throwing, a suicide, and several robberies brought about for the sake of this forty-grain weight
- Peak 3: index=860, pos_norm=0.8245, words=86000-86300
  - Preview: at once. I could, of course, borrow so trifling a sum ten times over from my friends, but I much prefer to make it a matter of business and to carry out that business myself. In my position you can readily understand tha

---

# Source: outputs/excitement_indep_clustering/cluster_report.md

# indep_winsize_5 Clustering Report (Feature-Primary, MA(W) with W=5)

## Executive Verdict
- The **feature branch** is primary for interpretation: selected `kmeans` with `k=5`.
- Feature quality at selected solution: silhouette `0.260`, Davies-Bouldin `0.888`, Calinski-Harabasz `8.08`.
- KMeans seed-stability is high (mean pairwise ARI `0.893`).
- DTW validation selected `k=2` with silhouette `0.085`; agreement with feature clusters is limited (ARI `0.059`, NMI `0.241`).

## Model Selection Diagnostics
- Elbow diagnostics (`k=1..10`) show largest inertia drop at `k=2` (drop `196.705`).
- Final model-selection policy remains silhouette-first over `k=2..6` for comparability with prior runs.
- DTW branch is used as trajectory-shape validation rather than the primary cluster definition.

## Feature Cluster Interpretation
- Cluster 1 (n=5): top signatures `mean_early (-1.11); p90_y (-1.11); mean_ma5 (-0.98)`; representative `1260 | Jane Eyre`.
- Cluster 2 (n=6): top signatures `prop_label_2 (+1.08); prop_label_0 (-0.95); p10_y (+0.94)`; representative `768 | Wuthering Heights`.
- Cluster 3 (n=6): top signatures `std_diff (+1.25); jump_ge_2_rate (+1.16); mean_abs_diff (+1.15)`; representative `1661 | The Adventures of Sherlock Holmes`.
- Cluster 4 (n=1): top signatures `prop_label_4 (+3.62); mean_late (+3.21); slope_position (+2.58)`; representative `1513 | Romeo and Juliet`.
- Cluster 5 (n=2): top signatures `max_y (-3.00); range_y (-3.00); lag1_autocorr (-1.84)`; representative `11 | Alice's Adventures in Wonderland`.

## Genre Composition by Cluster
- Cluster 1 is most concentrated in `Adventure` (2 books, 40.00% of cluster).
- Cluster 2 is most concentrated in `Adventure` (3 books, 50.00% of cluster).
- Cluster 3 is most concentrated in `Sci-Fi` (3 books, 50.00% of cluster).
- Cluster 4 is most concentrated in `Tragedy` (1 books, 100.00% of cluster).
- Cluster 5 is most concentrated in `Children's Fiction` (1 books, 50.00% of cluster).

## Feature vs DTW Agreement
- ARI: `0.059`
- NMI: `0.241`
- Interpretation: DTW captures broad trajectory shape, but does not strongly reproduce the feature-space grouping in this run.

```
## Figure Gallery (Embedded)
### Feature PCA scatter (legend-checked)
![Feature PCA scatter (legend-checked)](figures/feature_pca_scatter_feature_clusters.png)

### KMeans elbow diagnostics
![KMeans elbow diagnostics](figures/feature_elbow_kmeans_inertia.png)

### Feature k-sweep quality metrics
![Feature k-sweep quality metrics](figures/feature_k_sweep_quality_metrics.png)

### DTW k-sweep silhouette
![DTW k-sweep silhouette](figures/dtw_k_sweep_silhouette.png)

### Genre counts by feature cluster
![Genre counts by feature cluster](figures/genre_by_feature_cluster_counts.png)

### Genre proportions by feature cluster
![Genre proportions by feature cluster](figures/genre_by_feature_cluster_proportions.png)

### Top feature signatures heatmap
![Top feature signatures heatmap](figures/cluster_feature_signature_heatmap_top12.png)

### Feature vs DTW contingency heatmap
![Feature vs DTW contingency heatmap](figures/cluster_method_contingency_heatmap.png)

### Feature cluster member trajectories (MA5)
![Feature cluster member trajectories (MA5)](figures/feature_cluster_member_trajectories_ma5.png)

### Feature/DTW centroid trajectories (raw)
![Feature/DTW centroid trajectories (raw)](figures/cluster_centroid_trajectories_raw.png)

### Feature/DTW centroid trajectories (MA5)
![Feature/DTW centroid trajectories (MA5)](figures/cluster_centroid_trajectories_ma5.png)

### DTW distance heatmap
![DTW distance heatmap](figures/dtw_distance_heatmap.png)

### DTW dendrogram
![DTW dendrogram](figures/dtw_dendrogram.png)

### Cluster size comparison
![Cluster size comparison](figures/cluster_size_comparison.png)

## Practical Use / Caveats / Next Steps
- Use now: interpretable archetype grouping from feature clusters and MA5 pacing profiles.
- Use with caution: DTW-driven hard grouping, because selected DTW split is coarse in this run.
- Caveat: small sample size (20 books) means cluster boundaries should be treated as exploratory.
1. Add bootstrap re-sampling of books to quantify cluster stability confidence intervals.
2. Add constrained model selection that penalizes singleton clusters for DTW branch.
3. Evaluate agreement with external weak labels (genre/twist ranks) for external
```

validity.

# Source: outputs/excitement_indep_clustering/insights.md

# indep_winsize_5 Clustering Insights (MA(W), current run W=5)

## 1) Setup and Integrity
- Books analyzed: **20**
- Input checks passed: label existence, shape normalization, range `[0,4]`, integer-like, and embedding length alignment.
- Feature set size: **33** numeric features per book.

## 2) Model Selection Results
- Feature branch selected: `kmeans` with `k=5` (silhouette=0.260, DB=0.888, CH=8.1).
- KMeans stability (mean pairwise ARI across seeds): 0.893.
- DTW branch selected: `average-linkage` with `k=2` (silhouette=0.085).

## 3) Cross-Method Agreement
- Adjusted Rand Index (feature vs DTW): **0.059**
- Normalized Mutual Information: **0.241**
- Agreement table: `outputs/excitement_indep_clustering/tables/cluster_method_agreement.csv`.

## 4) Feature-Branch Cluster Archetypes
- Cluster 1 (n=5): mean_early (delta_z=-1.11); p90_y (delta_z=-1.11); mean_ma5 (delta_z=-0.98). Representative: 1260 | Jane Eyre.
- Cluster 2 (n=6): prop_label_2 (delta_z=+1.08); prop_label_0 (delta_z=-0.95); p10_y (delta_z=+0.94). Representative: 768 | Wuthering Heights.
- Cluster 3 (n=6): std_diff (delta_z=+1.25); jump_ge_2_rate (delta_z=+1.16); mean_abs_diff (delta_z=+1.15). Representative: 1661 | The Adventures of Sherlock Holmes.
- Cluster 4 (n=1): prop_label_4 (delta_z=+3.62); mean_late (delta_z=+3.21); slope_position (delta_z=+2.58). Representative: 1513 | Romeo and Juliet.
- Cluster 5 (n=2): max_y (delta_z=-3.00); range_y (delta_z=-3.00); lag1_autocorr (delta_z=-1.84). Representative: 11 | Alice's Adventures in Wonderland.

## 5) DTW-Branch Shape Archetypes
- Cluster 1 (n=19): prop_label_4 (delta_z=-0.19); mean_late (delta_z=-0.17); slope_position (delta_z=-0.14). DTW medoid: 11 | Alice's Adventures in Wonderland.
- Cluster 2 (n=1): prop_label_4 (delta_z=+3.62); mean_late (delta_z=+3.21); slope_position (delta_z=+2.58). DTW medoid: 1513 | Romeo and Juliet.

## 6) Practical Reading
- Use feature clusters as primary interpretable archetypes.
- Use DTW clusters as trajectory-shape validation; disagreement flags uncertain archetypes.
- MA5 centroid plots are better for coarse pacing patterns than raw per-chunk spikes.

## 7) Next Steps
1. Add bootstrap stability over book-resampling for selected k.
2. Compare clustering using raw vs MA5-only feature subsets.
3. Add external validation against genre/twist metadata as weak labels.

## Provenance
- Figures: `outputs/excitement_indep_clustering/figures/*.png`
- Tables: `outputs/excitement_indep_clustering/tables/*.csv`

# Source: outputs/excitement_linear/interpretation.md

# Verdict
- Chunk-level precision verdict: **not strong enough yet** for reliable absolute excitement scoring.
- Trend-level interpretation verdict (smoothed): **usable as a coarse proxy** for excitement dynamics across a novel.
- Bottom line: the model captures direction and pacing shape after smoothing, but misses too many raw chunk-level fluctuations to be used as a precise per-chunk label replacer.
- Smoothing notation in this report uses **MA(W)** where `W` is notebook-configured (`MA_WINDOW`). For the current run, `W=9` per `outputs/excitement_linear/tables/presentation_mae.csv`.

# Evidence by Figure
1. `outputs/excitement_linear/figures/train_loss_curve.png`
- Train and test MSE both drop quickly then plateau.
- Small train/test gap supports a generalization-stable but capacity-limited model.

2. `outputs/excitement_linear/figures/prediction_scatter_train_test.png`
- There is diagonal trend (signal exists), but the vertical spread is wide.
- This indicates moderate association, not high-fidelity point prediction.

3. `outputs/excitement_linear/figures/residual_hist_train_test.png`
- Residual spread is broad on both splits.
- Test residual mean is slightly positive, indicating slight overprediction on test novels.

4. `outputs/excitement_linear/figures/novel_overlay_test_16.png`
5. `outputs/excitement_linear/figures/novel_overlay_test_113.png`
6. `outputs/excitement_linear/figures/novel_overlay_test_768.png`
7. `outputs/excitement_linear/figures/novel_overlay_test_1342.png`
- Raw curves mismatch frequently at chunk level.
- MA(W) curves align better in direction and local pacing, but spikes and sharp transitions are often damped or missed (current run: `W=9`).

8. `outputs/excitement_linear/figures/mae_raw_vs_moving_average.png`
- Smoothing substantially reduces MAE for both train and test.
- This reinforces that the model is more useful for trend interpretation than raw pointwise estimation.

# Quantitative Summary
- Source: `outputs/excitement_linear/tables/global_metrics.csv`
  - Test RMSE: **1.193**
  - Test MAE: **1.041**
  - Test R2: **0.136**
  - Train RMSE: **1.203**
  - Train MAE: **1.064**
  - Train R2: **0.173**
- Interpretation: train/test are close, so underfitting is more likely than overfitting.

- Source: `outputs/excitement_linear/tables/presentation_mae.csv` (`ma_window` is runtime-configurable and equals `9` in the current run)
  - Train MAE raw: **1.064** vs MA(W): **0.328** (`W=9`)
  - Test MAE raw: **1.041** vs MA(W): **0.349** (`W=9`)

- Additional diagnostics from current outputs and model weights:

  - Raw correlation is moderate: test `corr(y, y_hat) ≈ 0.38`.
  - Smoothed correlation is stronger: test `corr(MA(W, y), MA(W, y_hat)) ≈ 0.70` (current run: `W=9`).
  - This supports trend-level use more than chunk-level use.

- Test-novel coverage from `outputs/excitement_linear/tables/per_novel_metrics.csv`
  - 16 | Peter Pan: RMSE `1.191`, MAE `1.033`, R2 `0.086`
  - 113 | The Secret Garden: RMSE `1.136`, MAE `0.968`, R2 `0.168`
  - 768 | Wuthering Heights: RMSE `1.234`, MAE `1.077`, R2 `0.112`
  - 1342 | Pride and Prejudice: RMSE `1.192`, MAE `1.057`, R2 `0.053`

# What the Model Can and Cannot Be Used For

Use now	Do not use yet
Relative trend profiling across chunks	Chunk-level absolute excitement scoring
Chapter/segment pacing summaries	Fine-grained spike detection
Cross-book comparison using smoothed trajectories	Threshold-based high-excitement decisions (e.g., alert when >=3)

# Risks / Caveats

- Label imbalance is significant. Label `4` is rare (~`1.7%`), limiting extreme-excitement calibration.
- The model is a single linear map, so nonlinear semantics and context interactions are under-modeled.
- Smoothing improves interpretability but can hide abrupt local events.
- Current conclusions are based on one fixed 16/4 novel split; split variability is not yet reported.

# Next Experiments (Prioritized)

1. Add stronger linear baselines on the same split.
- Try Ridge and ElasticNet against current linear GD baseline.

2. Use objective functions that better match the target.
- Try weighted MSE or ordinal-aware losses to handle rare high labels.

3. Add lightweight temporal features.
- Add `x_t - x_{t-1}` and short rolling embedding context before the linear head.

4. Keep strict acceptance criteria for adoption.
- Accept an upgrade only if both conditions hold:
- Test RMSE improves materially.
- Worst-case test-novel RMSE improves.

5. If linear upgrades saturate, then test small nonlinear capacity.
- Try a compact MLP with early stopping and same novel-level split protocol.

# Provenance

- Figures: `outputs/excitement_linear/figures/*.png`
- Tables: `outputs/excitement_linear/tables/global_metrics.csv`, `outputs/excitement_linear/tables/per_novel_metrics.csv`, `outputs/excitement_linear/tables/presentation_mae.csv`
- Split: `outputs/excitement_linear/tables/split_manifest.csv`

# Source: outputs/excitement_variant_analysis/insights.md

# Excitement Variant Analysis Insights (MA(W), current run W=5)

## 1) Dataset and Integrity Summary

- Books analyzed: **20** (reused split: 16 train / 4 test novels).
- All three label variants passed shape, range, and alignment checks (`len(label)==T`, labels in `[0,4]`, integer-like).
- `winsize_5` block-constancy check passed (no within-block variation for 5-chunk blocks).

## 2) Variant Comparison Findings

- Test split global metrics by variant (lower RMSE/MAE better; higher R2/corr better):
  - `indep_winsize_5`: RMSE=0.749, MAE=0.613, R2=0.075, corr=0.306, MA(5) MAE=0.270
  - `winsize_5`: RMSE=0.989, MAE=0.768, R2=0.152, corr=0.411, MA(5) MAE=0.621
  - `base`: RMSE=1.193, MAE=1.041, R2=0.136, corr=0.380, MA(5) MAE=0.460

- Pairwise label agreement indicates the variants are materially different sources, not trivial rewrites.
- See: `tables/variant_pairwise_agreement_global.csv` and `tables/variant_pairwise_agreement_per_book.csv`.

## 3) indep_winsize_5 Verdict (Primary Focus)

- Chunk-level reliability: **limited/moderate**. Raw pointwise error remains substantial.
- Trend-level utility (MA(5)): **useful as a coarse proxy** for trajectory/pacing interpretation.
- indep global train: RMSE=0.727, MAE=0.593, R2=0.193, corr=0.440, MA(5) MAE=0.244.
- indep global test: RMSE=0.749, MAE=0.613, R2=0.075, corr=0.306, MA(5) MAE=0.270.

## 4) Book-Level Highlights (indep_winsize_5, test novels)

- 1342 | Pride and Prejudice: RMSE=0.726, MAE=0.588, R2=0.043, corr=0.263, MA(5) MAE=0.260.
- 768 | Wuthering Heights: RMSE=0.747, MAE=0.615, R2=0.088, corr=0.327, MA(5) MAE=0.264.
- 113 | The Secret Garden: RMSE=0.747, MAE=0.611, R2=-0.002, corr=0.205, MA(5) MAE=0.272.
- 16 | Peter Pan: RMSE=0.816, MAE=0.678, R2=0.076, corr=0.283, MA(5) MAE=0.309.

## 5) Use Now vs Avoid Now (indep_winsize_5)

- Use now: relative trend profiling, chapter-level pacing summaries, cross-book smoothed trajectory comparison.
- Avoid now: chunk-level absolute scoring, spike-triggered threshold decisions, fine-grained event detection from raw predictions.

## 6) Next Experiments + Acceptance Criteria

1. Compare Ridge/ElasticNet against current linear GD head on the same split.
2. Add temporal features (`x_t - x_{t-1}`, short rolling context) while keeping linear head.
3. Try imbalance-aware objectives for rare labels.
4. Accept a new model only if both improve: test RMSE and worst-case test-novel RMSE.

## Provenance

- Figures: `outputs/excitement_variant_analysis/figures/*.png`
- Tables: `outputs/excitement_variant_analysis/tables/*.csv`
- Models: `outputs/excitement_variant_analysis/model/*.npz`

```
# Source: outputs/pca_analysis/insights.md

# PCA Component Insights (Global + Book)

## Corpus-Level Component Diagnostics
- PC1 explained_variance_ratio=0.0719 | cumulative=0.0719
- PC2 explained_variance_ratio=0.0650 | cumulative=0.1369
- PC3 explained_variance_ratio=0.0385 | cumulative=0.1754
- PC4 explained_variance_ratio=0.0312 | cumulative=0.2066
- PC5 explained_variance_ratio=0.0300 | cumulative=0.2366

## Component Semantics (Exemplar-Based)
- PC1:
  + positive exemplar: 1184 | The Count of Monte Cristo | score=0.526 | shore of
 the Mediterranean. If Bonaparte landed at Naples, the whole coalition would be
on foot before he could even reach Piombino; if he l
  + positive exemplar: 1184 | The Count of Monte Cristo | score=0.524 | my honor
,’ replied M. de Villefort; ‘they fancy that their countryman is still emperor.
You have mistaken the time, you should have told me
  - negative exemplar: 16 | Peter Pan | score=-0.447 | up.” “Were not the leaves
 at the foot of the window, mother?” It was quite true; the leaves had been foun
d very near the window. Mrs. Darlin
  - negative exemplar: 16 | Peter Pan | score=-0.432 | or the only available tre
e is an odd shape, Peter does some things to you, and after that you fit. Once y
ou fit, great care must be taken to
- PC2:
  + positive exemplar: 1342 | Pride and Prejudice | score=0.513 | I go on I shal
l displease you by saying what I think of persons you esteem. Stop me, whilst yo
u can.” “You persist, then, in supposing his s
  + positive exemplar: 1342 | Pride and Prejudice | score=0.506 | and whose asto
nishment at being so addressed was very evident. Her cousin prefaced his speech
with a solemn bow, and though she could not he
  - negative exemplar: 36 | The War of the Worlds | score=-0.552 | the end. It w
as dropping off in flakes and raining down upon the sand. A large piece suddenly
 came off and fell with a sharp noise that brou
  - negative exemplar: 521 | Robinson Crusoe | score=-0.542 | a W. and by S. sun
, or thereabouts, which, in those countries, is near the setting. Before I set u
p my tent I drew a half-circle before the
- PC3:
  + positive exemplar: 1661 | The Adventures of Sherlock Holmes | score=0.482 |
“S. H. for J. O.” Then he sealed it and addressed it to “Captain James Calhoun,
Barque _Lone Star_, Savannah, Georgia.” “That will await him
  + positive exemplar: 103 | Around the World in Eighty Days | score=0.452 | cap
tain forgot in an instant his anger, his imprisonment, and all his grudges again
st his passenger. The “Henrietta” was twenty years old; i
  - negative exemplar: 84 | Frankenstein; Or, The Modern Prometheus | score=-0.4
10 | of my dead mother in my arms; a shroud enveloped her form, and I saw the gr
ave-worms crawling in the folds of the flannel. I started from my
  - negative exemplar: 345 | Dracula | score=-0.409 | he may baffle us for years
; and in the meantime!--the thought is too horrible, I dare not think of it even
 now. This I know: that if ever th
- PC4:
  + positive exemplar: 521 | Robinson Crusoe | score=0.482 | yet come to the pit
ch of hardness to which it has since, reproached me with the contempt of advice,
 and the breach of my duty to God and my
  + positive exemplar: 521 | Robinson Crusoe | score=0.453 | myself; and if I sh
ould not fall into their hands, what I should do for provision, or whither I sho
uld bend my course; none of these thought
  - negative exemplar: 175 | The Phantom of the Opera | score=-0.422 | must reac
h Christine at all costs. He therefore went on his knees also and hung from the
trap with both hands. "Let go!" said a voice. And h
  - negative exemplar: 35 | The Time Machine | score=-0.414 | hurry on ahead!” “
```

```
To discover a society,” said I, “erected on a strictly communistic basis.” “Of a
ll the wild extravagant theories!” began t
- PC5:
  + positive exemplar: 55 | The Wonderful Wizard of Oz | score=0.401 | Dorothy,
clapping her hands. “Oh, let us start for the Emerald City tomorrow!” This they
decided to do. The next day they called the Winkies
  + positive exemplar: 55 | The Wonderful Wizard of Oz | score=0.373 | a winged
laugh; “but as we have a long journey before us, I will pass the time by telling
 you about it, if you wish.” “I shall be glad to he
  - negative exemplar: 1661 | The Adventures of Sherlock Holmes | score=-0.445 |
 the body was eventually recovered. It proved to be that of a young gentleman wh
ose name, as it appears from an envelope which was found in h
  - negative exemplar: 1661 | The Adventures of Sherlock Holmes | score=-0.442 |
 beast. His cry brought back his son; but I had gained the cover of the wood, th
ough I was forced to go back to fetch the cloak which I had d

## Book-Level Highlights
Highest PCA trajectory volatility (mean_speed):
- 345 | Dracula | mean_speed=0.1426 | p95_speed=0.2523
- 84 | Frankenstein; Or, The Modern Prometheus | mean_speed=0.1421 | p95_speed=0
.2623
- 175 | The Phantom of the Opera | mean_speed=0.1362 | p95_speed=0.2490

Strongest PCA-speed / acceleration coupling (k=7):
- 768 | Wuthering Heights | corr_speed_a=0.4642 | corr_speed_s=0.3766
- 1342 | Pride and Prejudice | corr_speed_a=0.3879 | corr_speed_s=0.3471
- 1513 | Romeo and Juliet | corr_speed_a=0.3427 | corr_speed_s=0.3857

Most atypical temporal component trends (|corr(PC, position)|):
- 43 | The Strange Case of Dr. Jekyll and Mr. Hyde | PC4 corr=0.5632 | q=0.0014
- 1513 | Romeo and Juliet | PC5 corr=-0.5549 | q=0.0015
- 175 | The Phantom of the Opera | PC2 corr=-0.5012 | q=0.0022

## Sensitivity Across k = 5, 7, 11
- k=5: median corr_speed_s=0.3819, median corr_speed_a=0.2358
- k=7: median corr_speed_s=0.3214, median corr_speed_a=0.2597
- k=11: median corr_speed_s=0.2544, median corr_speed_a=0.2670

## Caveats
- PCA components are derived from embedding geometry and require semantic triang
ulation with text exemplars.
- Association metrics are correlational and do not establish causal narrative me
chanisms.
- Missing/invalid signal artifacts are skipped and logged in integrity tables.



---

# Source: outputs/summary.md

# Story Trajectory Pipeline Summary

## Created artifacts
- Books processed: 20
- Feature rows: 60
- k values: [7, 5, 11]
- PCA dimensions saved per book: [2, 5]
- KMeans clusters file: /Users/kongfha/Desktop/Time_Series_Mining/story-trajecto
ry-analysis/outputs/clusters_kmeans.csv
- Hierarchical clusters file: /Users/kongfha/Desktop/Time_Series_Mining/story-tr
ajectory-analysis/outputs/clusters_hier.csv
```

```
- DTW distance matrix: /Users/kongfha/Desktop/Time_Series_Mining/story-trajectory-analysis/outputs/dtw_distance_k7.npy
- DTW resample length: 200
- Metadata file: /Users/kongfha/Desktop/Time_Series_Mining/story-trajectory-analysis/data/metadata.csv (exists=True)
- Sanity plot: /Users/kongfha/Desktop/Time_Series_Mining/story-trajectory-analysis/outputs/sanity_signal_examples.png


---
```