Cleaning & Preprocessing

- Removed duplicates at multiple stages to ensure consistency and remove redundancy.
- Replaced stringified empties (e.g., '[]', 'None', '') across key metadata fields with actual NaN.
- Dropped columns not used for modeling or redundant after transformation:
- images_review, bought_together, subtitle, author, details, average_rating, rating_number, videos, asin, features, description, images_meta, timestamp.
- Converted timestamp (int64 in ms) to UTC datetime: timestamp_utc.
- Text Standardization:
    - Trimmed and lowercased all textual fields (e.g., title_review, text, title_meta, store).
    - Preserved structure in list-like columns (e.g., features, description, images_meta, categories) while applying cleaning.
    - ID Standardization: Trimmed and cleaned asin, parent_asin, and user_id for consistency.
    - Type Casting: Cast rating column to int after coercing invalid values to 0.
    - Null Summary: Missing value analysis performed and tracked across both df and top_5_df.

Feature Engineering

Category Extraction:

- Parsed categories column from stringified list format.
- Created hierarchical levels: cat_0 to cat_4, and category_depth as the number of category levels.
- Dropped cat_0 due to high null rate and limited variance.
- Imputed missing values in cat_3 and cat_4 using fallback from cat_2 and cat_3 respectively.

Flags & Indicators:

- verified_purchase_flag: Binary flag derived from verified_purchase.
- helpful_vote_clipped: Applied upper limit of 5 for training robustness in BPR-style models. • Log Transformation:
- price_log: Log-transformed price to address skew in distribution

Textual Enrichment:

- features_clean and description_clean: Flattened and cleaned list-style text for SBERT-friendly embeddings.

Image Metadata Breakdown: Transformed images_meta into 3 structured fields:

- main_image_url → Best image (priority: hi_res > large > thumb)
- num_images → Count of image entries
- hi_res_images → Full list of hi-res (or large fallback) image URLs

Aggregated Rating:

- avg_rating_parent: Mean rating at parent_asin level to smooth variant-level noise. Rounded to 1dp for consistency.