

Rosa

Glossary of report and manifest terms

This glossary defines every term and field that appears in Rosa's two evidence outputs: the **PDF report** (produced for Diagnose and Training jobs) and the **Run Manifest** (the machine-readable JSON record emitted for every job). Your PDF report and your Run Manifest are two views of the same measured figures, so they never disagree on a value.

1. The bias measurement (read this first)

Rosa measures bias by asking a single question: **how well can a model predict the protected attribute from the rest of your data?** If the other columns carry no information about it, a predictor can only do as well as chance. If they carry information (directly, or through proxy columns), a predictor beats chance, and that margin above chance is the bias.

Five terms do all the work:

Term	What it means
Average Discriminator Performance	The held-out predictive accuracy a trained detector network reaches when predicting the protected attribute from the rest of the data. Higher means the attribute is more recoverable, i.e. more bias. Reported as an average over several training runs, with a standard deviation.
Random Baseline Accuracy	The accuracy that same measurement reaches by chance on data of this shape. It is established by re-running the measurement on randomly shuffled copies of the protected attribute: with the labels shuffled there is no real signal, so whatever accuracy the detector still reaches is pure chance. This is the honest floor the bias score is measured against.
Best Random Accuracy	The accuracy of always guessing the most common group (for example, if 58% of rows are "Male", always guessing "Male" scores 0.58). It is used only to set the scale of the score.
Bias	How far the detector beats chance, expressed on a 0 to 100% scale. This is the "before" score.
Residual Bias	The same Bias measurement, re-run on your data after Rosa has debiased it. This is the "after" score. Training jobs only.

1.1 How the bias score is calculated

The score uses **both** chance measures. It subtracts the **Random Baseline Accuracy** (the honest chance floor), then rescales onto a 0 to 100% axis using the **Best Random Accuracy**:

$$\text{Bias} = (\text{Average Discriminator Performance} - \text{Random Baseline Accuracy}) / (100\% - \text{Best Random Accuracy})$$

- The numerator is how far the detector beats the chance floor (the Random Baseline Accuracy).
- The denominator is the room a perfect predictor has above the Best Random Accuracy, so the score reads 0% when the detector does no better than chance and 100% when the attribute can be predicted perfectly.

The two baselines are usually close, so for a quick read you can treat them as one number. Worked example, with both around 60%: if the detector reaches 80%, the bias is $(80 - 60) / (100 - 60) = 50\%$, halfway to a perfect predictor. Residual Bias uses the identical formula on the debiased data.

Your Run Manifest records both baselines separately: `null_accuracy` (the Random Baseline Accuracy, the shuffled-label chance level that is subtracted) and `best_random_accuracy` (the Best Random Accuracy, the most-common-group rate that sets the scale).

1.2 Significance figures (the permutation test)

Because the chance level is measured as a distribution over many shuffled runs, Rosa also reports whether the bias it found is statistically significant rather than noise:

Term (manifest field)	What it means
Null accuracy (<code>null_accuracy</code>)	The mean accuracy across the shuffled-label runs (the chance floor above).
Null accuracy, 95th percentile (<code>null_accuracy_p95</code>)	The 95th percentile of that chance distribution. A detector accuracy above this is above what chance produces 95% of the time, i.e. a significance threshold.
Bias p-value (<code>bias_p_value</code>)	The probability of seeing an accuracy this high if there were no real bias. A small value (for example 0.05 or below) means the detected bias is unlikely to be chance.
Bias detected (<code>bias_detected</code>)	A true/false flag: is the bias statistically significant?

Each has a residual (after-debiasing) counterpart on training jobs: `residual_null_accuracy`, `residual_null_accuracy_p95`, `residual_bias_p_value`, `residual_bias_detected`.

2. Run Manifest fields

The Run Manifest is written once per job, cannot be changed afterwards, and is retained as your permanent evidence record. It never contains your row-level dataset, the debiased output, or the trained model. Job modes are `diagnose`, `remove_bias_training`, and `remove_bias_inference`.

2.1 Top-level fields

Field	Definition
<code>job_id</code>	Unique identifier for the job.
<code>workspace_id</code>	The workspace (API key) the job belongs to.

Field	Definition
timestamp_submitted / timestamp_started / timestamp_completed	Job lifecycle timestamps.
total_processing_seconds	Wall-clock processing time.
mode	diagnose, remove_bias_training, or remove_bias_inference.
row_count	Rows processed (after sampling, if any).
sampled_from_row_count	The original uploaded row count when your dataset was down-sampled for training; null if no sampling occurred.
num_variables	Number of input features before one-hot encoding.
bias_columns	The protected attribute(s) you submitted.
declared_columns	The column roles you set: {bias_columns, ignore_columns, cat_columns}. Makes the manifest self-describing about how the run was scoped.
input_hash	SHA-256 of your raw input CSV bytes, computed before any processing, formatted "sha256:<64 hex>". Lets you recompute the hash from your original file and prove the exact file was processed unaltered.
schema_hash	SHA-256 of your column names.
config_hash	SHA-256 of the job configuration.
container_digest	Identifier of the Rosa build that ran the job.
rosa_version	The Rosa version that ran the job.
model_format_version	Format version of the trained model; null for diagnose.
output_artifact_hashes	A SHA-256 hash for each deliverable file, so you can verify a download has not been altered.
idempotency_key	An optional key you supply so a retried submission returns the same job instead of creating a duplicate; null if not supplied.
job_status	complete or failed.
error	The reason a job failed, when applicable.
engine_attempts	The number of processing attempts the job took; 1 for almost all jobs.
bias_summary	The bias measurement block (see 2.2).
debiasing_adjustments	Per-column debiasing detail, diagnose jobs only (see 2.3).

Field	Definition
feature_associations	The proxy view (see 2.3).
feature_associations_post	Proxy view recomputed on the debiased output, training jobs only (see 2.3).
column_stats	Per-column statistics (see 2.4).

2.2 The bias_summary block

Field	Definition
best_random_accuracy	The most-common-group rate (scale term, see section 1).
average_discriminator_performance (+_std)	Detector accuracy on the original data, with its standard deviation.
residual_discriminator_performance (+_std)	Detector accuracy on the debiased data (training jobs).
bias	The "before" bias score.
residual_bias	The "after" bias score (training jobs).
null_accuracy (+_p95)	The shuffled-label chance level and its 95th percentile.
bias_p_value / bias_detected	Significance of the "before" measurement (see 1.2).
residual_null_accuracy (+_p95), residual_bias_p_value, residual_bias_detected	The same significance figures for the "after" measurement (training jobs).

2.3 Feature-level fields

Field	Definition
debiasing_adjustments	For a Diagnose job, each column's share of the total adjustment Rosa would apply when debiasing (the values sum to 1.0). This is not each column's correlation with the protected attribute: Rosa adjusts all features together, so this is not a ranking of which features carry the bias. To see which features track the protected attribute, use feature_associations.
feature_associations	The proxy view: a ranked list giving each feature's own association with the protected attribute on a 0 to 1 scale (point-biserial, correlation-ratio, or Cramer's V depending on the column type). It is a transparent single-feature check you can reproduce yourself. It complements, and does not replace, Rosa's full analysis, which also detects and removes non-linear and combined proxies that a single-feature score cannot see. Present on diagnose and training jobs.

Field	Definition
feature_associations_post	The same association view recomputed on the debiased output, so you can see the proxy signal collapse after debiasing. Training jobs only.

2.4 The column_stats block (per column)

Field	Definition
column_name	The column name.
datatype	The column's data type.
missing_values	Count of missing values.
categorical	Whether the column is treated as categorical.
num_categories / categories / counts	For categorical columns: the number of distinct values, the values themselves, and their row counts.
bias_contribution	The column's share of the debiasing adjustment (see debiasing_adjustments).
min / max / std_dev	For numeric columns: summary statistics.
min_change_pct / max_change_pct / std_dev_change_pct / num_categories_change_pct	The percentage change in each statistic between your original and debiased data, evidencing that Rosa preserves each column's distribution bit-for-bit, to float64 CSV precision.

Because column_stats can include category values from your data, and the manifest is retained permanently, place direct identifiers (names, email addresses, customer IDs) in ignore_columns rather than cat_columns so they are never recorded.

3. PDF report sections

The Diagnose report and the Training report share most sections; differences are noted.

- **Processing Summary.** The run's facts: dataset name and size, whether it was sampled, feature counts before and after one-hot encoding, start and end times, and the input and output files.
- **Bias Summary (Diagnose) / Bias Removal Summary (Training).** The headline table (see section 1). The Diagnose report shows Bias Characteristic, Random Baseline Accuracy, Average Discriminator Performance, and Bias. The Training report adds a "Bias (before debiasing)" figure and a Residual Bias figure, and its Average Discriminator Performance is measured after debiasing. A dial (gauge) renders the score visually.
- **Feature Associations (Diagnose).** A ranked table of each feature's own association with the protected attribute, a reproducible check that complements Rosa's full analysis.
- **Debiasing Adjustment by Feature (Diagnose).** A ranked table of how much Rosa would adjust each feature during debiasing. This is not a ranking of which features predict the protected attribute.
- **Run Evidence.** Makes the report a standalone audit artifact, in five parts: (1) dataset and sampling disclosure; (2) integrity and provenance hashes, identical to the manifest, so the report and the manifest

agree byte for byte; (3) the declared column scope; (4) protected-group representation (counts and shares per group); (5) a scope-and-limitations statement.

- **Dataset Statistics.** Per-column statistics, including the change percentages that show your data's distribution is preserved bit-for-bit, to float64 CSV precision.
- **Appendix.** Plain-English definitions of the terms used in the report.

The PDF report is meant to be read by a person, so any section that lists individual features (Feature Associations, Debiasing Adjustment by Feature, and the per-column Dataset Statistics) shows at most the 100 most important features for that section, ranked by importance within the section, to keep the report a manageable length. The Run Manifest is machine-readable, so it is never capped: it records every feature. Where a section reaches the limit, the report says so (for example "top 100 of 240") and points you to the manifest for the complete list.

4. The same term across your report, manifest, and the portal

Concept	Report label	Manifest field	Portal label
Protected attribute	Bias Characteristic	<code>bias_columns / declared_columns.bias_columns</code>	Bias characteristic
Chance floor	Random Baseline Accuracy	<code>null_accuracy (+ _p95)</code>	Random Baseline Accuracy
Detector accuracy (before)	Average Discriminator Performance	<code>average_discriminator_performance (+ _std)</code>	Average Discriminator Performance
Detector accuracy (after)	Average Discriminator Performance (after debiasing)	<code>residual_discriminator_performance (+ _std)</code>	Average Discriminator Performance
Before bias score	Bias	<code>bias</code>	Bias
After bias score	Residual Bias	<code>residual_bias</code>	Residual bias
Proxy strength per feature	Feature Associations	<code>feature_associations / feature_associations_post</code>	Feature Associations
Debiasing adjustment per feature	Debiasing Adjustment by Feature	<code>debiasing_adjustments</code>	Debiasing adjustment by feature