ELYREON DATASET DOCTOR

Dataset Verification & Quality Audit Report • Job ID: **demo_03_conversation_validation_splits**

1. Executive Summary & Record Statistics

Metric	Value	Metric	Value
Total Raw Records	15	Valid Cleaned Records	13
Modified Records	1	Rejected Records	2
Exact Duplicates	0	Near Duplicates	0
Syntax Errors	0	PII Incidents Detected	0

2. Token Distribution & Length Characteristics

Total Tokens	Min Tokens	Mean Tokens	Median Tokens	95th Percentile
460	22	35.38	36	45

3. Partition Isolation & Data Leakage Audit

Audit Result: PASSED: Zero leakage detected between train and test/val partitions.
Train records: 10 | Val records: 1 | Test records: 2

4. Verification Methodology & Scientific Integrity

Validation applies strict deterministic parsing, character-level Unicode NFC normalization, 64-bit SimHash Hamming-distance near-duplicate detection, and RFC-compliant PII scanning. **Notice:** ELYREON Dataset Doctor does not fabricate synthetic quality improvements or guarantee downstream model benchmark performance without empirical training runs.

5. Cryptographic Manifest Checksums

Artifact	SHA-256 Checksum
Input Original File	`729b89529d176d731f84eb9c94220eee6ad4e26da066403b31f12f8028a3ecba`
Cleaned Dataset	`6a1e3b78403b90177330e9f6a483d4c1fdb7d4d6a45db543087afd6ce6788fe7`
Rejected Records	`baae86988b23850ec9bb21e2fd0ab63912a90ed2fcfd37302cd4e0bf2a3ae434`
Change Log	`87b76df7958bc931646b09c20face62e2bb234362b6aeea9349012a3c71df84d`