ELYREON DATASET DOCTOR

Dataset Verification & Quality Audit Report • Job ID: **demo_01_finetuning_cleanup**

1. Executive Summary & Record Statistics

Metric	Value	Metric	Value
Total Raw Records	12	Valid Cleaned Records	9
Modified Records	2	Rejected Records	3
Exact Duplicates	1	Near Duplicates	1
Syntax Errors	2	PII Incidents Detected	0

2. Token Distribution & Length Characteristics

Total Tokens	Min Tokens	Mean Tokens	Median Tokens	95th Percentile
240	14	26.67	28	38

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	`6492f1b52ab2d18cd9e4349664c9d7f3cacb822d07f0701b1ca277e7fe821b47`
Cleaned Dataset	`418579b007a4e9298003ffff7efe1d2dc02daa8add14c1e4d63bdf8079d5d8f6`
Rejected Records	`9724011b35a3ed2680b1dfa6ac025cca5247a0405ba2a57cc8161aa148366993`
Change Log	`3fb4de43f06744f41dec7cbd8bc50399f91124b2241bb7972de0b9f927e884f1`