

Cutterhead Test Data Generation

Realistic synthetic eDiscovery datasets — built from a detailed scenario bible, ready for any eDiscovery platform.

Each dataset is a complete fictional eDiscovery matter — the right mix of signal, noise, privilege, and red herrings to train reviewers, benchmark analytics, and realistically demonstrate software.

THE SCENARIO BIBLE

- Scenario Overview**
Case type, legal theory, cast of characters, and key events
- Company & World**
Org chart, third parties, code language, and email domains
- Custodian Profiles**
5–8 personas: writing style, platform habits, and behavioral red flags
- Timeline**
Every event anchored — keeps documents internally consistent
- Document Type Specs**
Format rules for email, Teams, Slack, attachments, and loose files
- Signal & Noise Targets**
Count plan with ratios for signal, noise, privilege, and red herrings

SUPPORTED SCENARIOS

- Product Liability & Concealment
- Trade Secret Theft
- Price Fixing / Antitrust
- Securities Fraud
- Employment & HR Matters
- Data Privacy / DSARS
- Insurance Fraud
- Government Contracts / False Claims
- Money Laundering
- Environmental Liability
- Merger & Acquisition
- Workplace Investigations

Don't see your use case? Just ask.

THE DELIVERABLE

5,000 – 10,000 documents per dataset

- ✓ **Metadata**
Date, Subject, From, To, CC, BCC on every record
- ✓ **Extracted Text**
Full searchable text for all document types
- ✓ **Native Files**
EML, DOCX, PPTX, XLSX, TXT, RSMF
- ✓ **Document Types**
Emails, attachments, and loose files
- ✓ **Threaded Email**
Thread IDs and complete reply chains
- ✓ **Bible Mapping**
Every record mapped to its bible section
- ✓ **Standard DAT Files**
Easy to load into any eDiscovery platform

CONSTRAINTS — Every dataset is grounded in reality yet entirely fictional. All companies, custodians, law firms, and events are invented. No real people, corporations, or legal proceedings are referenced.

IDEAL FOR TESTING

✓ **Standard eDiscovery Workflows**
Ingestion, search, review coding, issue & privilege tagging, and productions

✓ **Legal AI Assistants**
Ground truth for RAG, summarization, and document Q&A

✓ **LLMs & Frontier Models**
Fine-tuning, benchmarking, and hallucination evaluation corpora

01 **Define the Scenario**
Case type, legal theory, and timeline locked in the bible

02 **Build the World**
Company, custodians, code language, and noise profiles

03 **Generate Content**
AI generates and checks every document guided by the bible

04 **Deliver**
Load file, native files, extracted text, bible mapping

Don't see exactly what you need? Just ask.
Every dataset is custom-built. If a scenario, format, or platform isn't listed, we can build it.