What a Million Epstein Documents Actually Contain

A Corpus-Scale Computational Audit of the Complete DOJ Release

RRecktek LLC

July 4, 2026

What a Million Epstein Documents Actually Contain

A Corpus-Scale Computational Audit of the Complete DOJ Release

RRecktek LLC

July 4, 2026

Table of Contents

Abstract

The Epstein Files Transparency Act (Public Law 119-38) obliged the U.S. Department of Justice to publish, in a searchable and downloadable format, its investigative files from the Jeffrey Epstein and Ghislaine Maxwell prosecutions. Between December 2025 and January 2026 the DOJ released twelve datasets. This work takes the complete release — 1,378,652 documents — and treats it as a measurement problem rather than a reading problem. Every document was ingested into a PostgreSQL corpus, given a full-text index and a 768-dimension semantic embedding, and every image-bearing document was additionally passed through a vision-language model that recorded document type, setting, visible and redacted faces, and explicit-content scores. This report counts 46 subject terms and 48 named individuals across the whole corpus, and measures the imagery directly. The headline findings are quantitative: the corpus is overwhelmingly three bulk datasets; its subject matter is trafficking logistics, finance, and an academic-and-finance social network, not the occult; machine vision found and measured tens of thousands of faces, most of them redacted; explicit imagery is real but rare; and the loudest conspiracy claims about the files effectively vanish when counted at scale, while the documented crimes do not.

The Corpus

All figures in this report were verified directly against the corpus on 2026-07-04. The release is not evenly distributed. Of the 1,378,652 documents, 99.2% sit in the three January 2026 bulk datasets

— DS9 (emails and internal DOJ correspondence), DS10 (media seized from Epstein properties), and DS11 (financial ledgers and travel records). The seven December datasets that drew the first wave of coverage are, together, a rounding error on the total.

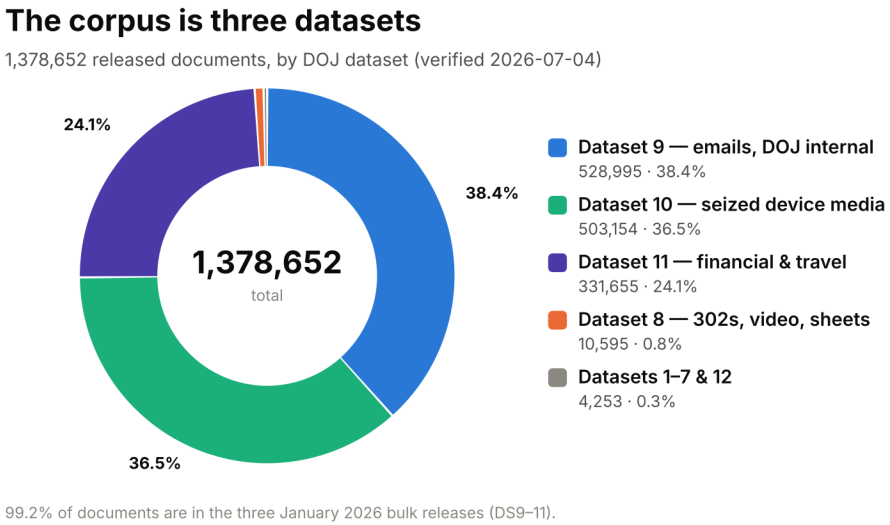

Figure 1: The corpus is three datasets. Composition of the 1,378,652 released documents by DOJ dataset.

Of the full corpus, 1,377,740 documents (99.9%) carry a semantic embedding, so the collection can be searched by meaning as well as by keyword. A keyword count tells you how often a term appears; an embedding search tells you what the nearest documents to an idea actually are, which is how several widely repeated claims were tested and found to be news clippings rather than evidence.

Method

Three independent measurements run over the corpus. **Full-text counting:** each document's extracted text is indexed as a PostgreSQL `tsvector`; topic and name magnitudes are exact counts of documents matching a search expression — a phrase for names, a term or boolean for topics. These counts describe presence, not guilt. **Semantic search:** every document is embedded as a 768-dimension vector, so querying with a concept returns the nearest documents by meaning — the test that separates a documented practice from a forwarded article that merely mentions it. **Machine vision:** the 74,462 image-bearing documents were each analyzed by a vision-language model that recorded, per page, the document type, setting, people count, faces visible and redacted, a facial-recognition-potential rating, and a nudity score; a separate on-premises classifier scored 73,424 rendered images for explicit content. No manual viewing of sensitive imagery was involved.

What the Files Are About: The Major Themes

Counting 46 subject terms across the corpus produces a clear hierarchy. The largest terms are structural — the physical makeup of an investigative file. The word *flight* appears in 65,564 documents (travel logistics and flight records), *Palm Beach* in 33,066 (the original crime scene and the 2005–2008 police investigation), and *trust* or *foundation* in 20,313 (the financial scaffolding). The investigative apparatus itself is heavily represented — *FBI* in 13,434 — and so is the human core of

the case: *victim* (13,395), *massage* (9,770, the recurring cover activity), *minor* (9,688), and *abuse* (8,639).

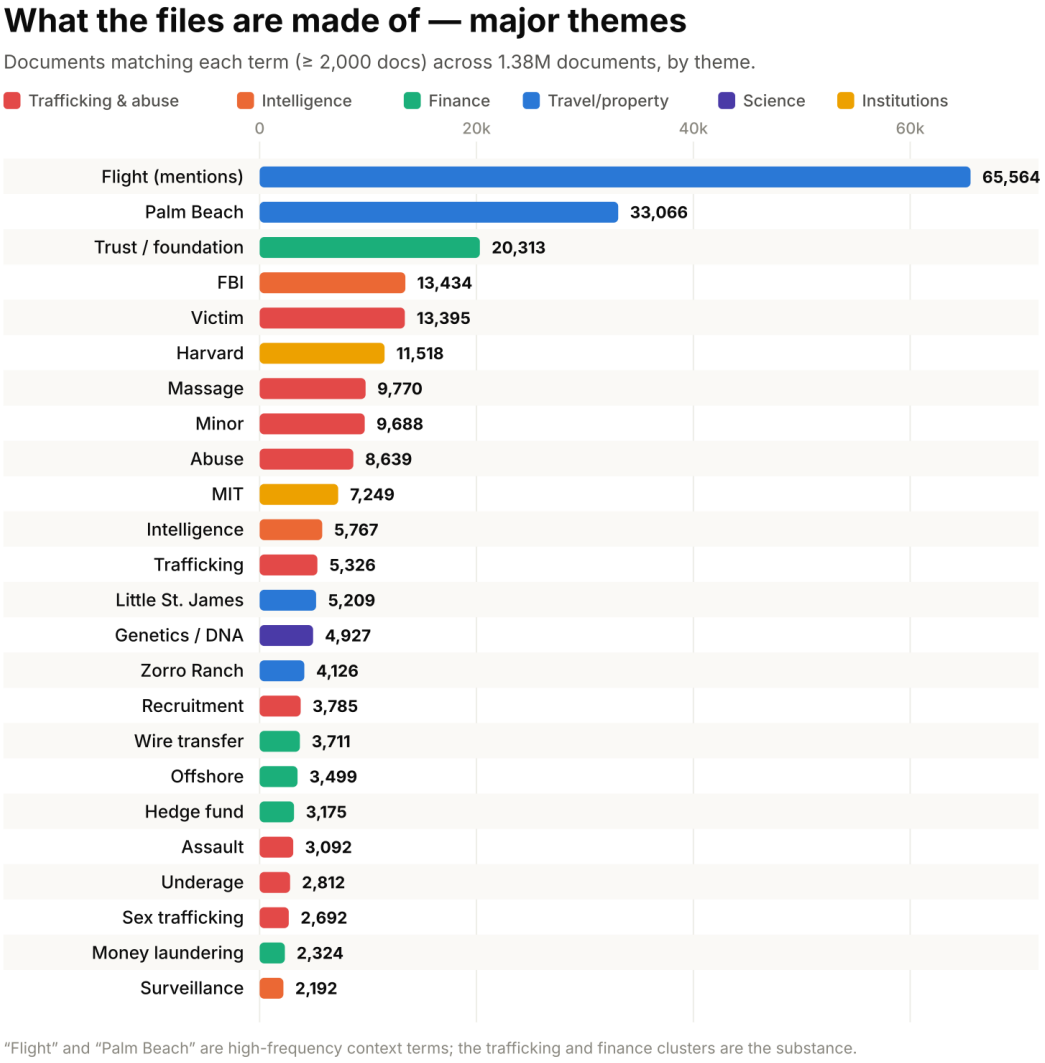

Figure 2: What the files are made of. Major topics (≥ 2,000 documents), colored by theme.

Two structural facts stand out. First, the academic network is quantitatively enormous — *Harvard* (11,518) and *MIT* (7,249) rank above most crime terms, a direct trace of Epstein's donations. Second, the trafficking and finance clusters are dense and mutually reinforcing: *trafficking* (5,326), *recruitment* (3,785), *assault* (3,092), *underage* (2,812), *sex trafficking* (2,692) on one side; *wire transfer* (3,711), *offshore* (3,499), *hedge fund* (3,175), *money laundering* (2,324) on the other. The named properties anchor the geography — *Little St. James* (5,209) and *Zorro Ranch* (4,126).

The Finer Topics, and the Conspiracy Tail

Below two thousand documents, the picture separates the merely specific from the essentially absent. Real but smaller clusters include the surveillance apparatus — *hidden camera* (441) and *blackmail*

(426) — the model agency *MC2* (1,033), and Epstein's documented scientific interests: *fertility* (789), *artificial intelligence* (749), and *eugenics* (536). *CIA* (1,924) leads this band but, read in context, sits mostly in bank-compliance and legal paperwork rather than any handler file.

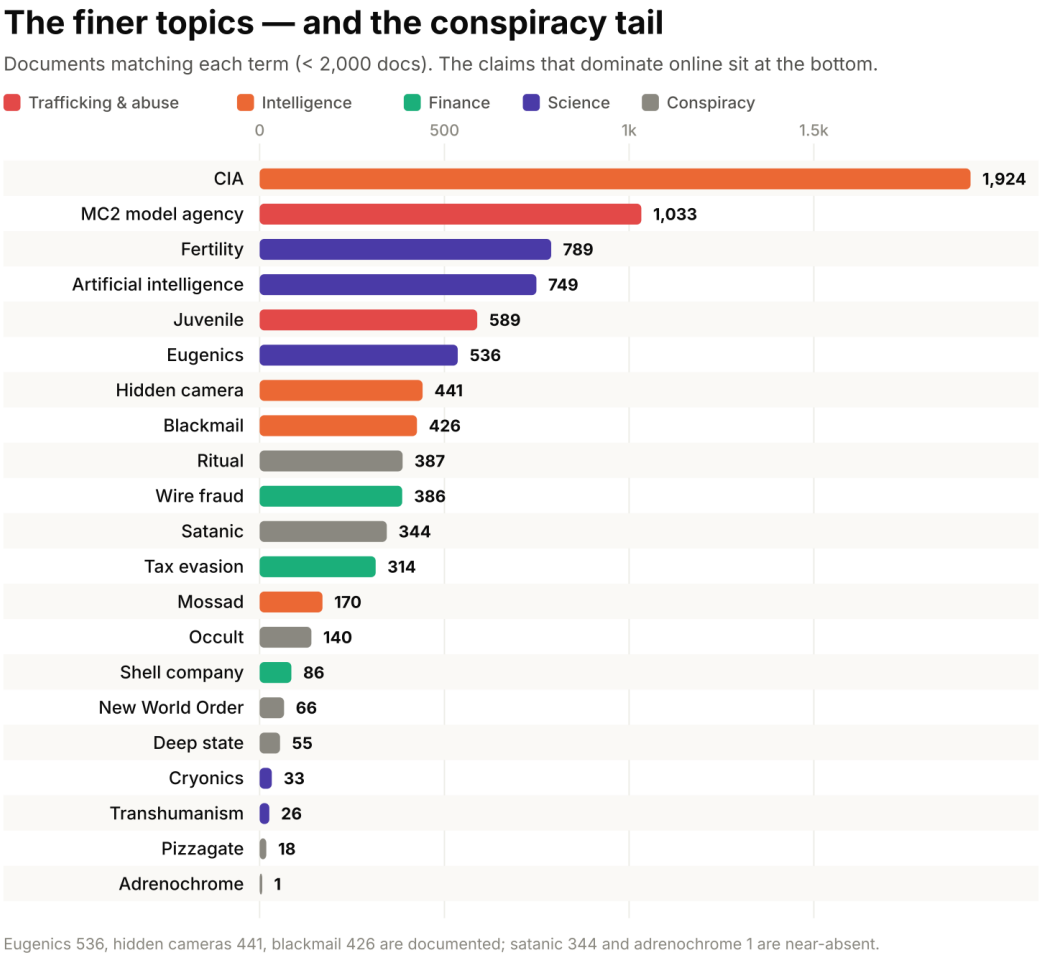

Figure 3: The finer topics and the conspiracy tail. Terms under 2,000 documents.

At the bottom of the chart are the claims with the loudest lives online. *Ritual* (387) and *satanic* (344) are almost entirely forwarded news articles; *occult* (140) doubles as a financial term meaning "hidden." *New World Order* (66), *deep state* (55), and *pizzagate* (18) are news commentary. *Adrenochrome* — the centerpiece of an entire conspiracy genre — appears in exactly one document across 1.38 million. The measurement does not merely fail to confirm these theories; it quantifies their near-total absence.

The Name Network

Forty-eight individuals were counted by exact phrase and grouped by role. The counts are mentions across the corpus and nothing more — presence in a document is not participation in a crime, and a large share of these mentions occur in news articles and FBI briefings that happened to be held in Epstein's own files.

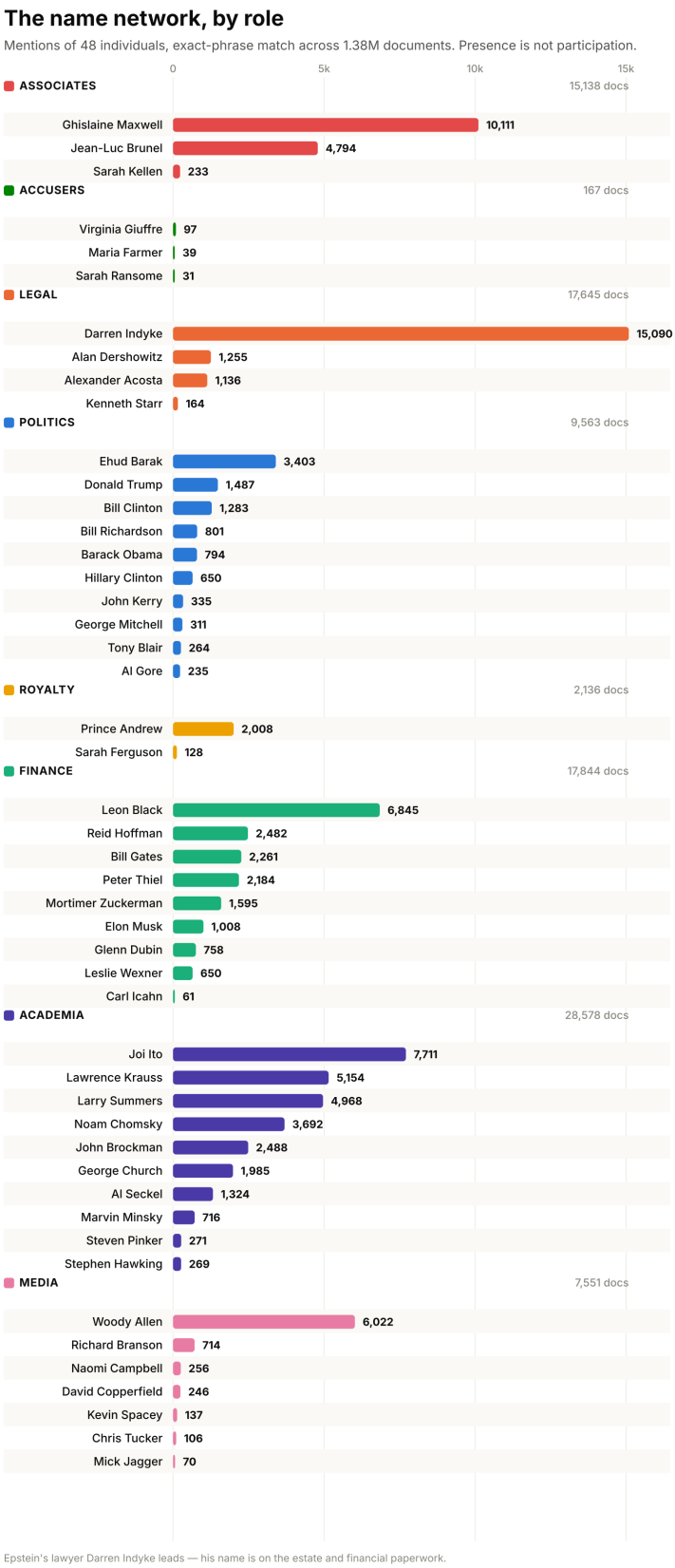

Figure 4: The name network by role. Mentions of 48 individuals across the corpus.

The ranking is led not by a politician or a celebrity but by the machinery of the estate. **Darren Indyke**, Epstein's personal attorney, tops the entire corpus at 15,090 documents because his name is on the financial and estate paperwork that makes up datasets 9 and 11. **Ghislaine Maxwell** follows at 10,111. The densest category is the academic-and-science network (over 28,000 documents across its names): **Joi Ito** (7,711, the MIT Media Lab director who resigned over Epstein's funding), **Lawrence Krauss** (5,154), **Larry Summers** (4,968), **Noam Chomsky** (3,692), **John Brockman** (2,488), and **George Church** (1,985). The finance category is comparably large — **Leon Black** (6,845), **Reid Hoffman** (2,482), **Bill Gates** (2,261), **Peter Thiel** (2,184), **Elon Musk** (1,008) — as is the political roster, led by former Israeli prime minister **Ehud Barak** (3,403), then **Donald Trump** (1,487) and **Bill Clinton** (1,283). **Prince Andrew** (2,008) dominates a small royalty category, and **Woody Allen** (6,022) leads media.

The most telling pattern is at the bottom. The women who brought the case — **Virginia Giuffre** (97), **Maria Farmer** (39), **Sarah Ransome** (31) — appear far less often than their centrality to the wrongdoing would suggest. This is a property of *this* corpus: the released files are dominated by financial records, correspondence, and seized media, not by victim testimony, which lived in the sealed civil litigation. Document frequency tracks administrative and correspondence volume, not moral weight.

The Imagery

More than half of the machine-classified images are photographs, and the settings the model assigned to them are mundane: exteriors, interiors, grounds, kitchens, a construction site, a beach. This is the visual signature of seizure and evidence photography — properties and their contents catalogued — rather than the imagery the files are imagined to contain.

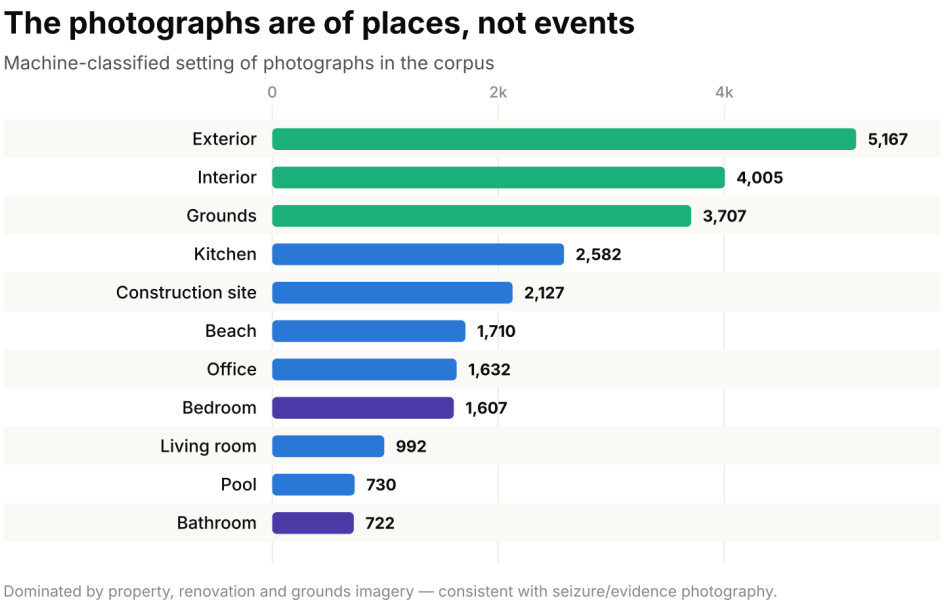

Figure 5: The photographs are of places, not events. Machine-classified setting of photographs.

Faces, and How Many Were Hidden

The vision pass counted faces on every image page, separating those left visible from those the DOJ had redacted. Across the imagery it found 56,699 faces. Nearly three-quarters of them — 40,747 — were redacted before release. This is the single clearest quantitative signature of the DOJ's disclosure posture: the release is heavily de-identified at the level of the individual face.

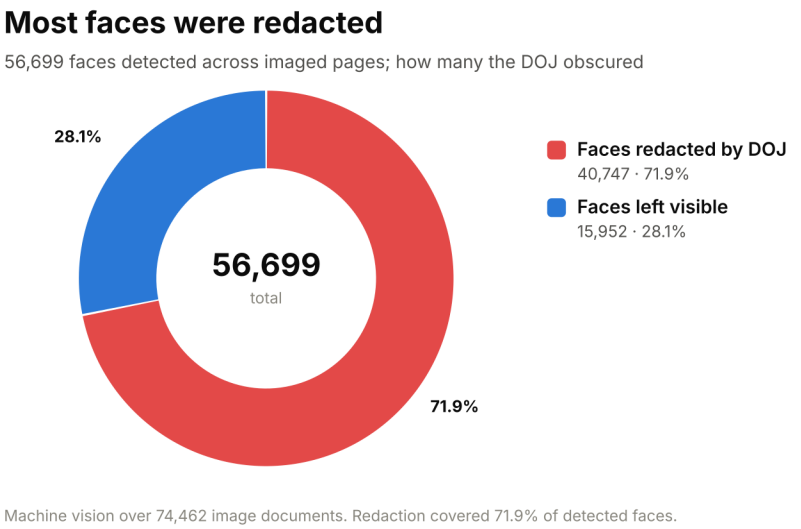

Figure 6: Most faces were redacted. Of 56,699 faces detected, 71.9% were obscured before release.

The model also rated each imaged page for facial-recognition potential — whether a clear, unredacted face is present that a recognition system could match. On that measure, 34,165 pages — about one in three imaged pages that were scored — are recognition-ready. The redaction is extensive, but it is not total.

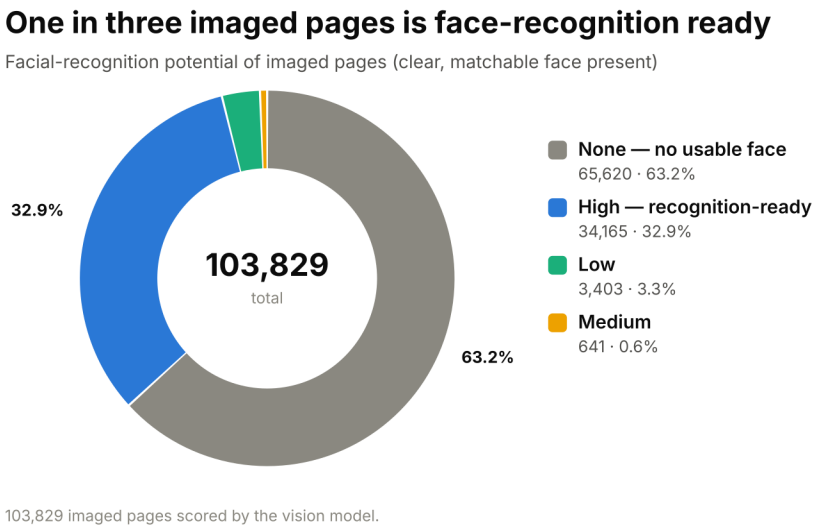

Figure 7: One in three imaged pages is face-recognition ready. Facial-recognition potential of imaged pages.

Explicit Content

Two independent methods measured nudity in the imagery, and they agree: explicit content is present but rare at corpus scale. A dedicated classifier scored 73,424 images; 71,540 of them fall in the lowest band, and only 126 images score at or above 0.50, with just 3 at or above 0.85. The vision-language model, working separately, flagged 1,003 pages as containing nudity, 166 as explicit, and 644 as containing minors.

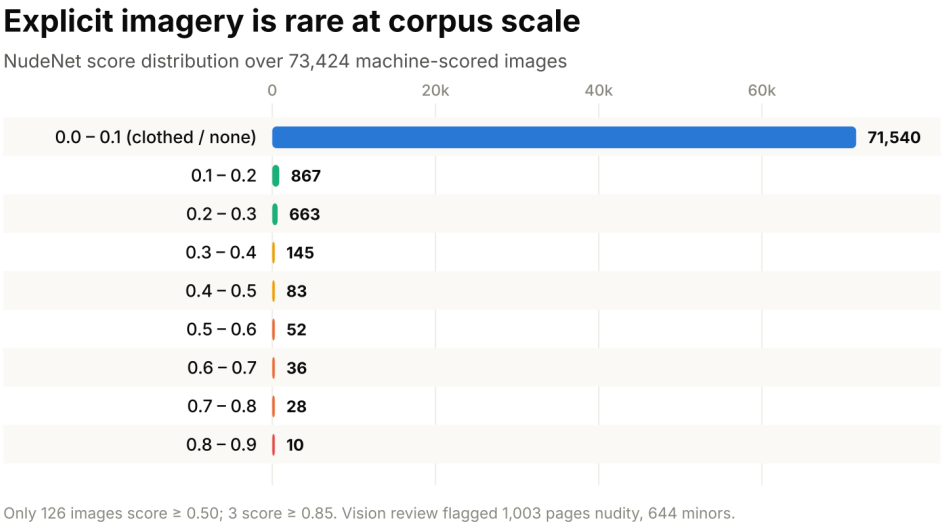

Figure 8: Explicit imagery is rare at corpus scale. NudeNet score distribution over 73,424 images.

The prevalence is low — the overwhelming majority of images are clothed, redacted, or non-explicit. But low prevalence across 1.38 million documents still resolves to hundreds of flagged pages, and the 644 pages the model associated with minors are exactly the material the law directs be handled with the most care. The value of counting is that it replaces speculation with a bounded, checkable figure.

Signal Versus Noise

The most useful thing a full-corpus count does is sort the documented from the imagined. **Documented with direct evidence:** the trafficking operation, the Brunel/MC2 recruitment pipeline, the financial machinery, and the surveillance of the properties are the substance of the files, present in FBI records, prosecution filings, correspondence, and financial documents. **Present but overstated:** the intelligence-operation and occult framings exist in the corpus almost entirely as forwarded news articles — Mossad (170), CIA (1,924, in bank-compliance and legal paperwork), satanic (344). **Effectively absent:** the claims with the loudest independent lives online do not survive contact with the corpus — adrenochrome appears once, semantic search for organized ritual returns a massage-service invoice and a magic-show invitation, and the UFO thread resolves to a film project in Epstein's correspondence.

What Is Not in the Release

A measurement of the released files is not a measurement of the whole case. The DOJ has stated it identified roughly six million pages of Epstein-related material and published about 3.5 million of them; the remainder is withheld under categories including child sexual abuse material, attorney-client privilege, and ongoing investigations. Independent integrity scans of the DOJ portal after publication reported that tens of thousands of documents — concentrated in DS9 — were removed from the servers in the weeks following release. The corpus analyzed here is the public release as captured, and it should be understood as a floor on what exists, not a ceiling.

Methods and Provenance

The corpus lives in PostgreSQL with the pgvector extension. Text was extracted with `pdftotext`; degraded scans were handled with a vision-language OCR pass rather than traditional OCR. Embeddings are 768-dimension vectors. Image analysis used a vision-language model for structured tagging and a separate on-premises classifier for explicit-content scoring. Every count in this report is reproducible as a single query against the indexed corpus, and all figures were re-verified on 2026-07-04 rather than carried forward from earlier runs. Counts describe what the documents contain; they do not, by themselves, establish what any named person did.