Clasio: Proof of Claims

# So, what exactly is Clasio? What does it do?

Clasio organizes your document chaos. Automatically. Snap, drop, or import a file and Clasio takes it from there.

### Capture without friction

* Upload PDFs, images, Word, Excel. Or import directly from Google Drive.
* Camera capture: take a photo of a document on mobile; it’s cleaned, classified, and filed.
* Voice mode: speak a request (e.g., “find my 2023 W‑2”) and naturally converse with Clasio

### Auto-organize intelligently

* Instant doc‑type detection (invoice, ID, contract), business category suggestions, topics, and concise 2–3 line summaries so you recognize files at a glance.
* Version history and bulk actions keep growing libraries tidy without duplicates.

### Find, fast: two modes

* Simple Search: quick, no‑nonsense search across names, content, tags, type, and date filters. This is for when you exactly know what document you need and that it exists.
* Ask Anything (AI Search): natural‑language questions with results explained via confidence indicators and a short “why this matched.” Toggle between modes anytime.

### Stay in control

* View AI insights for every document (summary, topic tags, classifications) and see search analytics that reveal which docs matched and why.
* Drive sync lets you import in bulk (5 docs at a time for now) and keep changes consistent both ways.

### Private by design

* Files live in your private cloud; access is only via short‑lived, signed URLs.
* Metadata (tags, folders, insights) is stored separately for auditability.
* Your documents are never used to train AI models.

### Production‑ready reliability

* Robust error handling, retries, and clean rollback keep uploads and downloads dependable.
* Clear “why this matched” explanations build trust and save time.

# Technical Specifications

### 1) Data privacy & tenancy (ingest → process → store → delete)

* Organization boundary with least‑privilege access: Clasio runs in an organization‑owned Google Cloud project (Cloud Identity). All data access uses service accounts with the minimum roles required; there is no human‑credential access path.
* User‑scoped storage paths: Deterministic, validated paths per user and document — users/{userId}/docs/{docId}/{originalFileName}; objects are encrypted at rest and fetched only via V4 signed URLs.
* Separation of concerns: Blobs in Google Cloud Storage; analysis artifacts and filing metadata in the database for clean auditing.
* Deletion you can trust: User‑initiated deletions cascade from metadata to storage within a defined window; we monitor and retry to ensure completion.

### 2) Model usage & training posture (Gemini 2.5 Flash‑Lite + Pro)

* No model training on customer data: Each request is private and not retained for training; we persist only compact analysis outputs required for product function (summary, topics, doc‑type, category).
* Right model for the job: Flash‑Lite for low‑latency extraction; Pro only for complex reasoning or long documents. All calls are server‑side, no client keys.

### 3) AI Search

* Semantic layer: multi‑field vector embeddings with cosine similarity and adaptive field weighting that picks the strongest signal rather than averaging weak ones.
* Lexical layer: precise keyword matching with field‑specific boosts and a content‑hierarchy weight; behavioral signals can lift exact matches users prefer.
* Quality layer: temporal freshness, content completeness, and user interaction history refine ordering, so the most useful document appears first.

### 4) Folders & sub‑folders (predictable + self‑healing)

* Human‑friendly hierarchy with normalized, collision‑safe paths; moves/renames are transactional across metadata and storage.
* AI‑suggested destinations (e.g., Taxes → 2024 → Returns) keep canonical paths consistent while allowing user overrides.

### 5) Performance & UX guardrails

* Background processing keeps uploads responsive; frequent queries are cached for speed.
* Confidence bars and concise “why‑match” explanations increase trust and reduce triage time.

### 6) Camera Capture & Voice Mode

* Camera capture (mobile): on‑device image cleanup (orientation, crop, glare reduction), progressive upload via signed URLs, instant classify‑and‑file on arrival.
* Voice mode: optional speech capture for hands‑free search or quick notes; audio is processed to text server‑side and not retained.
* Privacy holds: same signed‑URL pattern and least‑privilege access; nothing is exposed to the client beyond what’s necessary to complete the action.

## What we won’t do

* We don’t sell or share your personal data with advertisers.
* We don’t expose bucket credentials to the client; access is only via signed URLs.
* We don’t use your documents to train AI models. Ever.

## How you (or your security team) can verify us quickly

* Data‑use test: Upload a doc, delete it in‑product, then confirm the storage object and metadata are removed.
* Access test: Try an expired signed URL; it should return 403 (no public reads).
* Explainability test: Compare the same query in Simple vs. AI Search and review the confidence and “why‑match” indicators.