

TECHNICAL REPORT BD-TR-2026-01 * BASIS DESK RESEARCH DESK

Primary-Source Grounding and Automated Claim Verification in Real-Time Cryptocurrency Newsrooms

Architecture, Dual-Pass Audit Thresholds, and Empirical Evaluation of the Basis Desk Pipeline

Authors: Julian Mercer (ORCID: 0009-0003-1162-170X) and Basis Desk Research Desk

Published: 2026-10-01 | JEL: G14, G18, C81 | License: CC BY 4.0

Canonical HTML: <https://basisdesk.news/research/bd-tr-2026-01> | PDF: <https://basisdesk.news/papers/BasisDesk.pdf>

ABSTRACT

Digital asset markets suffer from severe information asymmetry, latency-driven rumor propagation, and unverified secondary reporting. This paper presents the architecture of Basis Desk (<https://basisdesk.news>), an autonomous, primary-source-first newsroom engineered to ingest, cluster, draft, and verify financial and cryptocurrency news in real time. Rather than rewriting secondary media coverage, the pipeline ingests documents directly from regulatory agencies, courts, corporate filings, protocol repositories, and public blockchains, using secondary media headlines strictly as event-discovery signals to locate underlying primary documents. Every candidate draft undergoes an independent second-pass grounding audit that verifies factual claims, dates, entities, and numerical figures against the source text and live market snapshots, enforcing a strict verbatim-overlap ceiling and a minimum publication quality score of 80/100.

Keywords: primary-source verification; automated fact-checking; cryptocurrency journalism; grounding audit; hallucination detection; financial NLP

1. Introduction and Problem Statement

Information cascades in cryptocurrency markets frequently originate from misinterpreted regulatory filings, spoofed announcements, or unverified social media posts. Traditional aggregation pipelines optimize for publication speed at the expense of provenance, amplifying factual errors across downstream syndication networks and large language model (LLM) retrieval systems.

Basis Desk addresses this failure mode by enforcing a strict separation between event discovery and factual grounding. Secondary media feeds are configured with `textUse=false`: their article bodies never enter a generation prompt. Instead, outbound links are parsed to retrieve the underlying primary document (e.g., an SEC Form 8-K, CFTC enforcement release, court docket, or protocol governance proposal) before any drafting workflow can execute.

2. Primary-Source Registry and Vector Clustering

The ingestion engine polls 70+ curated primary feeds across regulators (SEC, CFTC, DOJ, Federal Reserve, UK FCA, ESMA, ECB, MAS, SFC), SEC EDGAR material-event filings, protocol release registries, and security researchers at 5-to-30-minute intervals. Incoming documents are canonicalized, hashed (FNV-1a), and embedded into a 768-dimensional vector space (bge-base-en-v1.5) to cluster multi-source coverage of the same event within a 72-hour rolling window (cosine similarity ≥ 0.76).

To prevent re-reporting of already-covered events across rolling news cycles, a three-layer deduplication gate combines (i) in-tick entity token exclusion, (ii) a strongly-consistent 96-hour D1 lexical entity/number overlap check, and (iii) a 7-day Vectorize semantic similarity ceiling (0.855 at publish time).

3. Dual-Pass Grounding Audit and Anti-Copy Enforcement

Each eligible story cluster is processed by a durable multi-step workflow (gather -> draft -> verify -> repair -> publish). During the verification stage, an independent auditor model evaluates every factual assertion in the headline, deck, key points, At-a-Glance summary, and body against the numbered primary source texts and a timestamped market snapshot.

The deterministic verifier (verifyDraft) combines LLM audit findings with programmatic checks: citation indices [1..k] must map to valid ingested sources, verbatim 8-gram overlap with source texts must remain below threshold, advice or price-target phrasing triggers a hard fail, and drafts scoring below 80/100 undergo a targeted repair pass or are diverted to human editorial review.

4. Structured Output for Answer Engines and Researchers

Published articles expose structured At-a-Glance metadata (what happened, why it matters, who is affected, effective date, and regulatory action status), clean Markdown twins (<url>.md), Schema.org NewsArticle / AnalysisNewsArticle JSON-LD, and a public Model Context Protocol (MCP) server at <https://basisdesk.news/mcp>.

The longitudinal dataset of verified articles, source attributions, and audit metrics is released under CC BY 4.0 at <https://basisdesk.news/api/research/dataset.csv> and <https://basisdesk.news/api/research/dataset.jsonl>.

References and Live Reproducibility Endpoints

- [1] Basis Desk Editorial Policy & Verification Standards (2026).
<https://basisdesk.news/about/editorial-policy>
- [2] How Basis Desk Uses AI: Dual-Pass Grounding Architecture (2026).
<https://basisdesk.news/about/how-we-use-ai>
- [3] Basis Desk Longitudinal Dataset & Frictionless Data Package (2026).
<https://basisdesk.news/api/research/datapackage.json>
- [4] Basis Desk Model Context Protocol (MCP) Specification (2026).
<https://basisdesk.news/developers>

Appendix: Linked Basis Desk Verification & Data Endpoints

- * Basis Desk Home & Live Newsroom: <https://basisdesk.news/>
- * Research Portal & Longitudinal Datasets: <https://basisdesk.news/research>
- * Source-Verified Crypto News Feed: <https://basisdesk.news/news>
- * Research-Desk Deep-Dive Analysis: <https://basisdesk.news/analysis>
- * Regulator, Statute & Company Hubs: <https://basisdesk.news/entities>
- * Evergreen Crypto Explainers (300 Topics): <https://basisdesk.news/learn>
- * On-Chain Whale Tracker (\$5M+ Transfers): <https://basisdesk.news/data/whales>
- * Basis Desk Verify (Chrome & Firefox Extension): <https://basisdesk.news/tools/verify>
- * Smart-Contract Token Safety Checker: <https://basisdesk.news/tools/token-checker>
- * Model Context Protocol (MCP) Server for AI Agents: <https://basisdesk.news/developers>
- * Machine-Readable Dataset (CSV): <https://basisdesk.news/api/research/dataset.csv>
- * Machine-Readable Dataset (JSONL): <https://basisdesk.news/api/research/dataset.jsonl>
- * Frictionless Data Package (JSON): <https://basisdesk.news/api/research/datapackage.json>
- * BibTeX Citation Record: <https://basisdesk.news/api/research/cite.bib>