

TECHNICAL SYSTEM PAPER

Directed Herding Research Instrument

System Architecture, Automated Data Acquisition, Methodological Design, and
Research Contribution

Prepared by Fatima Aguilar

System audit date: July 18, 2026

System status. Production oriented research prototype with 10 validated institutional filing managers, automated weekly collection checks, a live retail-attention pilot, and descriptive signal generation. It is not yet a causal inference engine or a completed publication dataset.

Abstract

The Directed Herding Research Instrument is a web-based empirical research environment designed to study whether large institutional investment managers exhibit synchronized portfolio changes, whether those changes cluster around identifiable information events, and whether retail attention moves before or after the institutional pattern. The system separates data acquisition, validation, descriptive analytics, research governance, and user-facing exploration. Supabase provides the database, scheduled collection, server-side transformation, provenance, and read-only research views. The user interface was originally designed in Lovable, exported into an independently owned private GitHub repository, and is now deployed through Cloudflare Workers. As of the audit snapshot, the platform contains 10 validated Form 13F filing managers, 104,370 holdings records across 11,511 securities, 20 verified event records, 442 Wikimedia pageview observations, and 4,062 descriptive herding candidates divided into high, moderate, and exploratory tiers. The system is deliberately conservative: synchronized quarterly holdings are labeled candidate alignment signals, not proof of imitation, manipulation, or causation. This paper documents the system's purpose, data sources, automation schedule, institution-selection framework, analytical methodology, research contribution, governance controls, production deployment, limitations, and next empirical steps.

1. Research Objective and Design Philosophy

The platform operationalizes a research question that is difficult to answer with a single dataset: do major U.S. institutional managers move into or out of the same securities in a common direction, can those movements be associated with credible information events, and does retail attention follow, precede, or merely coincide with the institutional pattern? The website is designed to reduce the cost of moving from raw filings to reviewable hypotheses.

The central design principle is separation of evidence levels. Form 13F establishes reported quarter-end positions; it does not reveal transaction dates, motives, or complete economic exposure. Event feeds establish that information was released; they do not prove that a manager traded because of it. Pageviews measure attention; they do not measure orders. Accordingly, the system produces descriptive candidates and a structured review queue rather than causal conclusions.

Research safeguard. The labels “high,” “moderate,” and “exploratory” refer to descriptive synchronization thresholds inside the platform. They do not represent statistical significance or validated causal confidence.

2. How the Platform Advances the Research

The platform moves the project forward in six concrete ways:

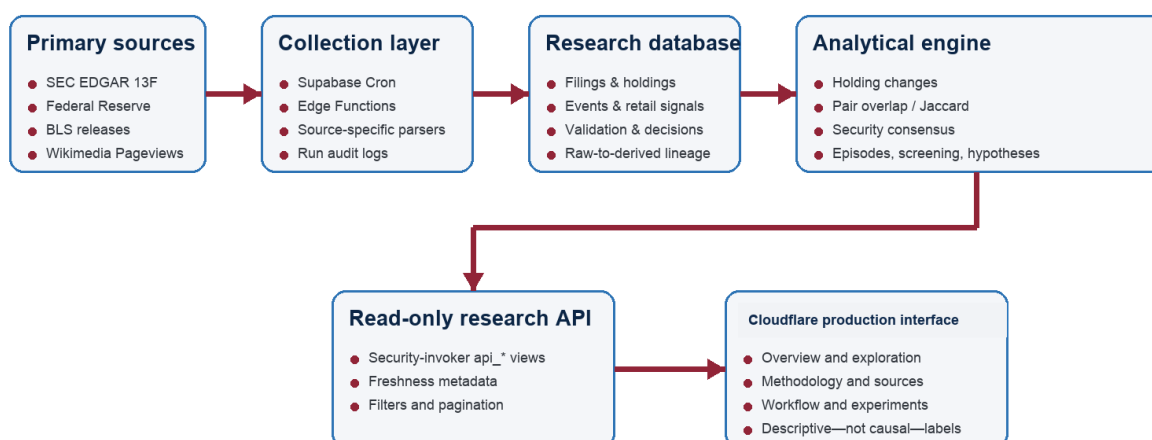
- It creates a reproducible institutional sample by linking parent brands to exact SEC filing manager names and CIKs.
- It turns heterogeneous 13F XML filings into normalized, comparable holdings and quarter-over-quarter change records.
- It surfaces cross-institutional overlap, synchronized increases, decreases, entries, and exits for targeted review.

- It creates a traceable event and retail attention context around candidate episodes.
- It records source choices, methodology decisions, validation status, experiments, open questions, and review actions in the same research environment.
- It provides an end to end path from raw public records to a hypothesis, alternative explanations, and a proposed empirical test.

This workflow is especially useful at the pre-estimation stage. It helps the researchers inspect entity scope, data continuity, mechanical explanations, and missing context before specifying formal models. It therefore reduces the risk of building sophisticated regressions on an invalid sample or an incorrectly interpreted filing.

3. Technical Architecture

Directed Herding Research Instrument — System Flow



Governance controls span every layer: source URLs, validation flags, methodology logs, run status, and explicit limitations.

Figure 1. Current system architecture and research-data flow. Source: system audit.

Supabase remains the research backend. PostgreSQL stores raw and normalized observations; Edge Functions acquire and parse external data; Supabase Cron invokes collection and analytical refresh functions; ingestion-run tables record success, failure, row counts, and error messages; and security-invoker API views expose read-only, research-ready outputs. The user interface was authored in Lovable and then exported as an independent TanStack Start/React codebase. The production application is now served by Cloudflare Workers, while the source of record is maintained in a private GitHub repository. The browser queries the Supabase api_* views with a publishable client key; it does not receive the service-role credential or perform the quantitative research calculations.

The Cloudflare-hosted user interface is organized into Core modules (Overview, Institutions, Holdings, Events, Retail Attention), Analysis modules (Herding Explorer, Institution Comparison, Daily Brief), and Research modules (Methodology, Data Sources, Journal Priorities, Research Workflow, Experiments). Each retained page has loading, empty, and error states and displays backend-derived freshness dates. Hosting

was changed without changing the research-data contract: the same Supabase views, validation rules, and provenance records continue to drive every page.

4. Data Sources and Automated Acquisition

4.1 Active production sources

Source	Research object	Access	Observation / refresh frequency	Status
SEC EDGAR Form 13F	Quarter-end institutional positions	Official EDGAR submissions/API	Quarterly data; checked weekly	Active for 10 filers
Federal Reserve Board	FOMC statements, minutes, speeches, testimony	Official releases/calendars	Event-driven source; collected weekly	Active
U.S. Bureau of Labor Statistics	CPI, PPI, employment, unemployment releases	Official release pages/calendar	Monthly/weekly source; collected weekly	Active
Wikimedia Pageviews	Firm-specific attention proxy	Official keyless REST API	Daily observations; refreshed weekly	Active pilot

SEC filings are the authoritative institutional-holdings source. Form 13F applies to managers meeting the statutory threshold and reports qualifying long positions as of calendar-quarter end; filings are generally due within 45 days after quarter end [1]. The weekly jobs therefore poll for new or amended filings, but weekly polling does not convert quarterly holdings into weekly holdings.

The event layer prioritizes official timestamps. Federal Reserve records distinguish policy statements from later minutes; BLS records distinguish release date from the economic reference period. These distinctions are necessary for event-window design and for preventing look-ahead bias.

The active retail pilot uses Wikimedia Pageviews because the API is public, official, stable, keyless, and reproducible [5]. It collects daily article-level views for up to 25 high-priority candidate securities with a 35-day lookback. The current run inserted 442 observations but completed partially because 12 article mappings or pageview requests failed. The system therefore exposes verification status and does not silently treat all mappings as valid.

4.2 Sources configured or cataloged but not currently scheduled

The research catalog also contains BEA macro releases, SEC enforcement and rulemaking events, Census data, Treasury releases, EIA energy data, issuer investor-relations materials, White House actions, researcher-coded geopolitical events, GDELT, Google Trends, WRDS 13F, TAQ retail classification, Robinhood, StockTwits, Reddit, and odd-lot volume. These entries document potential coverage and access decisions; they must not be described as live production feeds unless a working job, successful run, and populated output are present.

At the audit date, BEA and SEC-event ingestion jobs were marked enabled in configuration metadata but had no active cron entry or successful run. GDELT was a backup discovery layer. Google Trends remained in an older open methodology decision, but the active pilot had been superseded by Wikimedia Pageviews because unattended official API access and reproducibility were stronger. TAQ remains the preferred transaction-level retail measure if licensed access becomes available.

4.3 Actual automation schedule

Scheduled time	Job	Execution frequency	Function
Saturday 01:00 UTC	Federal Reserve event ingest	Weekly	Checks official Fed material
Saturday 01:10 UTC	BLS event ingest	Weekly	Checks official BLS releases
Saturday 01:20 UTC	Wikimedia retail attention	Weekly	Imports daily views; 35-day lookback
Sunday 02:00–02:45 UTC	Ten SEC 13F collection jobs	Weekly, staggered 5 min	Polls each validated filer
Sunday 04:00 UTC	Research-output refresh	Weekly	Recomputes comparisons, signals, episodes, and brief

The scheduler uses Supabase Cron/pg_cron to invoke database functions and authenticated Edge Functions. Supabase records job definitions in cron.job and run history in cron.job_run_details [6]. Collection is staggered to reduce simultaneous SEC requests. Source-specific code parses filings or feeds, preserves source URLs, upserts normalized records, and writes ingestion results with received, inserted, rejected, and error counts.

5. Institutional Sample Selection

The sample-selection framework was designed to prevent convenience sampling and cherry-picking. Six criteria are stored with explicit weights:

Criterion	Weight	Reason
Assets under management	25%	Scale and potential market influence
13F reporting continuity	20%	Consecutive quarters without gaps
Portfolio breadth	15%	Number and diversity of reportable positions
Active/passive research fit	15%	Ability to contrast discretionary and index-oriented behavior
Market relevance	15%	Visibility and plausible informational influence
Data quality	10%	Completeness, consistency, and filing-entity clarity

The final operational sample contains 10 exact filing managers whose SEC identities were verified and whose imported filings passed continuity checks. The set deliberately combines very large diversified and index-oriented managers with active managers and bank-affiliated institutions. This provides variation in business model and portfolio breadth while maintaining market relevance.

Research label	SEC filing manager	CIK	Periods	Latest loaded period
Goldman Sachs	GOLDMAN SACHS GROUP INC	886982	2	2026-03-31
Bank of America / Merrill	BANK OF AMERICA CORP /DE/	70858	2	2026-03-31
Morgan Stanley	MORGAN STANLEY	895421	2	2026-03-31
BlackRock	BlackRock, Inc.	2012383	2	2026-03-31
Vanguard	VANGUARD GROUP INC	102909	2	2025-12-31
State Street	STATE STREET CORP	93751	2	2026-03-31

Research label	SEC filing manager	CIK	Periods	Latest loaded period
Fidelity	FMR LLC	315066	2	2026-03-31
JPMorgan	JPMORGAN CHASE & CO	19617	2	2026-03-31
T. Rowe Price	PRICE T ROWE ASSOCIATES INC /MD/	80255	2	2026-03-31
Wellington	WELLINGTON MANAGEMENT GROUP LLP	902219	2	2026-03-31

Two governance caveats are material. First, Bank of America’s consolidated 13F is not equivalent to individual Merrill client portfolios. Second, Vanguard’s latest loaded report period was December 31, 2025 at the audit snapshot, while nine other managers had March 31, 2026. Analyses for March 2026 must therefore either exclude Vanguard, align all institutions to the common period, or explicitly model the unbalanced coverage.

The database defines the weighted criteria, but candidate weighted_total fields were not populated at the audit date. Consequently, the defensible current statement is that the top 10 were selected through the documented criteria plus verified filer identity and continuity—not that a completed numeric ranking mechanically produced the final ten. Populating and freezing those scores is a priority before publication.

6. Data Processing, Validation, and Provenance

- Entity resolution: parent brands are mapped to exact SEC filing-manager names and CIKs; ambiguous subsidiaries are not merged by brand name alone.
- Filing selection: the importer checks Form 13F-HR and amendments, preferring the research-relevant amended filing when appropriate.
- Normalization: filings, securities, holdings, voting authority, reported value, shares, and period metadata are stored in relational tables.
- Change construction: current and prior shares are compared to classify new, increased, decreased, exited, and unchanged positions.
- Validation: position-count continuity, period coverage, amendment handling, duplicate constraints, source URLs, and methodology_usable status are recorded.
- Auditability: every ingestion run can record status, timestamps, counts, error messages, and metadata; raw source references are retained.
- Frontend security: the Cloudflare-hosted application reads only security-invoker api_* views with a Supabase publishable key; mutation, service-role, and ingestion controls are not exposed to anonymous users.

7. Analytical Methodology

7.1 Pairwise portfolio comparison

For each pair of methodology-usable institutions, the system calculates common holdings, union holdings, simultaneous changes, Jaccard similarity, and the overlap coefficient. If A and B are the sets of securities held by two institutions, Jaccard similarity is $|A \cap B| / |A \cup B|$. The overlap coefficient is $|A \cap B| / \min(|A|, |B|)$. The second measure is useful when one portfolio is much broader than the other, while Jaccard penalizes the size of the union.

7.2 Security-level synchronization score

For each security and report period, the engine counts methodology-eligible institutions holding the security and institutions whose classified change is increased, decreased, new, or exited. The dominant direction is the unique maximum across those four categories; ties are labeled mixed. Let D be the number moving in the dominant direction, E the eligible-institution count, and C the total number with one of the four directional classifications. The platform computes:

$$\text{Participation component} = D / E$$

$$\text{Consensus component} = D / C$$

$$\text{Base signal strength} = (D / E) \times (D / C)$$

This score rewards both breadth across the institution universe and directional purity among observed changes. It is descriptive and does not correct for expected common buying, benchmark effects, or the security's baseline ownership prevalence.

7.3 Episode construction and prioritization

A security becomes a candidate episode when at least three institutions share a dominant increase, decrease, new-position, or exit classification. The research-priority score is:

$$\text{Priority score} = 0.55 \times \text{base signal strength} + 0.30 \times \text{breadth} + 0.10 \times \text{verified event context} + 0.05 \times \text{retail-attention context}.$$

Tier	Rule	Interpretation
High	$D \geq 6$ and consensus ≥ 0.80	Strong descriptive synchronization
Moderate	$D \geq 4$ and consensus ≥ 0.65	Material candidate requiring review
Exploratory	$D \geq 3$	Hypothesis-generating only
Insufficient	Below thresholds	Not promoted as an episode

For high and moderate episodes, the system generates a research hypothesis with the observation, a broad proposed mechanism, alternative explanations, and required tests. Alternatives explicitly include passive rebalancing, mechanical flows, reporting-entity changes, common fundamentals, and independent responses to the same public information.

7.4 Empirical methods prepared for the research program

Method	Role	Principal risk
LSV herding measure	Tests excess common buying/selling relative to an expected buy proportion	Benchmark definition and mechanical trading
Sias measure	Tests persistence and cross-sectional correlation in institutional demand	Common information can mimic herding
Event study	Estimates abnormal returns around precisely dated events	Confounding events and window choice
Panel fixed effects	Explains portfolio changes across institutions and quarters	Time-varying omitted variables; serial correlation
Difference-in-differences	Uses plausibly exogenous shocks and controls	Parallel-trends and contamination
Lead-lag analysis	Tests timing among events, institutional changes, and attention	Quarterly aggregation and spurious sequencing
Portfolio network analysis	Represents institutions as nodes and similarity as edges	Dense portfolios inflate overlap

The LSV and Sias approaches anchor the institutional-herding literature [7,8]. Event studies, fixed-effects panels, difference-in-differences, and lead-lag analysis are not yet estimated by the website; they are registered as planned empirical designs. This distinction is important: the existing engine ranks observations for research, while the formal statistical engine remains the next stage.

8. Retail-Attention Methodology

The active retail proxy is firm-level Wikipedia pageviews. For each mapped company article, the platform retains the daily view count, source URL, mapping status, a current seven-day average, a prior 28-day average and standard deviation, and an abnormal-attention z-score. The conceptual justification follows the investor-attention literature: search and information-seeking behavior can reveal attention even when it does not identify a trade [9]. Wikipedia usage has also been studied in relation to subsequent market movements [10].

Wikipedia was chosen for the active pilot over an unofficial Google Trends scraper because an official, keyless API improves unattended reproducibility and auditability. Google Search Volume remains a stronger conceptual benchmark in finance, while TAQ-classified retail order flow would be the preferred publication-grade behavioral measure if licensed access is available. Any paper must call Wikipedia an attention proxy, not a retail-trading dataset.

9. Current System Output and Readiness

Metric	Audit value	Interpretation
Validated institutions	10	All marked methodology-usable
Holdings records	104,370	Quarterly 13F observations
Unique securities	11,511	Ticker/sector enrichment incomplete
High-confidence descriptive signals	388	Threshold-based, not statistical significance
Moderate signals	1,805	Require review
Exploratory signals	1,869	Hypothesis generation
Review-ready episodes	388	Prepared for methodology review
Verified events	20	Current official-source coverage
Retail observations	442	Wikimedia pilot; partial first run
Average metadata completeness	33.3%	Material limitation

These counts make the website useful for research triage today. They do not make the dataset publication-ready. Only two consecutive filing periods are loaded for each institution, event history is shallow, retail coverage is a pilot, and security metadata completeness is approximately one third. Formal longitudinal inference requires additional quarters, enriched identifiers and sectors, stronger event coverage, and a validated retail-trading or attention series.

10. Reproducibility, Governance, and Research Ethics

- All source decisions record purpose, access method, frequency, cost or access barrier, quality concerns, legal or ethical notes, status, and methodology rationale.
- Methodology decisions store alternatives considered, rationale, reviewer risk, decision status, and whether Prof. Dalko's input is required.
- Official filing URLs and source URLs are preserved so a dashboard observation can be traced back to the originating record.
- Private or client-level Merrill and Goldman data are not assumed or represented; the active system uses public consolidated filing-manager data.
- Automated classifications are treated as candidates for human review, not ground truth.
- The website avoids claims of market manipulation and avoids inferring intent from public quarterly positions.

11. Limitations and Required Corrections Before Publication

- Quarterly temporal resolution: weekly collection checks do not reveal weekly trades. The platform cannot determine the exact order of institutional transactions from 13F alone.
- Disclosure delay and scope: 13F is delayed and excludes many shorts, derivatives, non-reportable securities, and intraperiod activity.
- Uneven period coverage: Vanguard currently lags the common March 2026 period.

- Source-catalog versus live-pipeline ambiguity: configured sources must be labeled separately from scheduled and successfully populated sources.
- Methodology-log drift: an older open decision still names Google Trends although Wikipedia is the active pilot; the decision record should be superseded or updated.
- Incomplete weighted ranking: selection criteria and weights exist, but numeric candidate totals are not populated.
- Metadata incompleteness: missing ticker and sector fields weaken sector analysis, event linking, and interpretation.
- Mechanical confounding: common benchmark exposure, corporate actions, identifier changes, and passive flows can resemble herding.
- Multiple testing and threshold calibration: thousands of candidates require false-discovery controls and out-of-sample threshold validation.

12. Recommended Research Roadmap

1. Freeze the institution-selection protocol: populate all weighted scores, document the candidate universe, and obtain Prof. Dalko's approval.
2. Backfill a multi-year quarterly panel and enforce common-period completeness rules.
3. Resolve security identifiers and enrich ticker, sector, industry, benchmark membership, and corporate-action history.
4. Activate and validate BEA and SEC-event pipelines; add issuer earnings and regulatory events with source-specific timestamps.
5. Repair failed Wikipedia article mappings, validate the abnormal-attention specification, and add Google SVI or TAQ as a robustness measure when access permits.
6. Estimate LSV and Sias benchmarks, then compare the platform's descriptive score with established herding measures.
7. Pre-register event windows, controls, exclusion rules, and multiple-testing corrections before causal analysis.
8. Export versioned analysis datasets and preserve the data, code, method, and source version behind every table and figure.

13. Production Deployment, Portability, and Ownership

The production frontend is no longer dependent on a Lovable publishing subscription. The exact Lovable project source was exported and committed to a private GitHub repository, github.com/fa366193/directed-herding-research-platform. Lovable remains part of the project history and may still be used as a visual authoring environment, but it is not required to serve the live research platform.

The exported application uses React and TanStack Start. Its dependency lockfile was retained, type checking and a production build were completed successfully, and Nitro generated a Cloudflare Module Worker plus 29 public assets. Wrangler uploaded the assets and Worker as one deployment unit. Cloudflare's static-assets model serves cacheable browser assets through its network while the Worker handles application requests [13,14].

The production deployment completed on July 18, 2026 (Eastern Time) as Cloudflare Worker version 58fd0f9a-2f58-4f0b-a975-f0ce375ece6f. The public research interface is available at <https://directed-herding-research-platform.directed-herding-research.workers.dev>. TLS provisioning completed successfully, and the endpoint was independently verified to return HTTP 200 over TLS 1.3. Supabase continues to run the collection, validation, analytical, and weekly scheduling layers; Cloudflare serves the frontend and does not replace the Supabase research backend.

The resulting ownership model is portable: GitHub is the source of record, Supabase is the research engine, and Cloudflare is the production delivery layer. Future releases should follow a controlled sequence of source update, type check, production build, Git commit, Cloudflare deployment, route-level verification, and recorded version identifier. The current workers.dev address is appropriate for this research-stage deployment; Cloudflare recommends a custom domain for business-critical production use [15].

14. Conclusion

The Directed Herding Research Instrument is best understood as research infrastructure: it standardizes public institutional holdings, official events, and retail-attention observations; validates and documents data choices; surfaces synchronized changes; and organizes those observations into reviewable hypotheses and empirical designs. Its strongest contribution is not an automated declaration that herding occurred. Its contribution is a reproducible chain of evidence that makes it easier for researchers to decide which patterns deserve formal testing, which alternative explanations must be ruled out, and which missing data constrain inference. With additional historical depth, event coverage, identifier enrichment, and publication-grade retail measures, the system can support a rigorous empirical directed-herding research program.

References

- [1] U.S. Securities and Exchange Commission. “Frequently Asked Questions About Form 13F.” Updated March 6, 2026. <https://www.sec.gov/rules-regulations/staff-guidance/division-investment-management-frequently-asked-questions/frequently-asked-questions-about-form-13f>
- [2] Federal Reserve Board. “FOMC Meeting Calendars, Statements, and Minutes.” <https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm>
- [3] U.S. Bureau of Labor Statistics. “Economic News Releases and Release Calendar.” <https://www.bls.gov/bls/newsrels.htm>
- [4] U.S. Bureau of Economic Analysis. “BEA Data and Release Information.” <https://www.bea.gov>
- [5] Wikimedia Foundation. “Wikimedia Pageviews API.” <https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews>
- [6] Supabase. “Cron: Schedule Recurring Jobs with Cron Syntax in Postgres.” <https://supabase.com/docs/guides/cron>
- [7] Lakonishok, J., Shleifer, A., and Vishny, R. W. (1992). “The Impact of Institutional Trading on Stock Prices.” *Journal of Financial Economics*, 32(1), 23–43.
- [8] Sias, R. W. (2004). “Institutional Herding.” *Review of Financial Studies*, 17(1), 165–206. <https://doi.org/10.1093/rfs/hhg035>
- [9] Da, Z., Engelberg, J., and Gao, P. (2011). “In Search of Attention.” *Journal of Finance*, 66(5), 1461–1499. <https://doi.org/10.1111/j.1540-6261.2011.01679.x>
- [10] Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., and Preis, T. (2013). “Quantifying Wikipedia Usage Patterns Before Stock Market Moves.” *Scientific Reports*, 3, 1801. <https://doi.org/10.1038/srep01801>
- [11] Directed Herding Supabase project. Live schema, cron configuration, Edge Functions, ingestion logs, methodology records, and read-only API views. System audit conducted July 18–19, 2026.

[12] Herding Insights Lovable project. Original interface design, route structure, Supabase query hooks, data-freshness handling, and research-interface organization. Source exported at Lovable commit ee927b791baa3c0c291430376f01d8d3f75f53cc on July 18, 2026.

[13] Cloudflare. “Static Assets.” Cloudflare Workers documentation. <https://developers.cloudflare.com/workers/static-assets/>

[14] Cloudflare. “Wrangler deploy.” Cloudflare Workers documentation. <https://developers.cloudflare.com/workers/wrangler/commands/workers/#deploy>

[15] Cloudflare. “Routes and domains.” Cloudflare Workers documentation. <https://developers.cloudflare.com/workers/configuration/routing/>

[16] Directed Herding Research Platform source repository. Private GitHub repository, github.com/fa366193/directed-herding-research-platform. Production migration commits 6c4fbc1, 006e41f, and dc49ca9; audited July 18, 2026.

[17] Directed Herding Research Platform production deployment. Cloudflare Worker directed-herding-research-platform, version 58fd0f9a-2f58-4f0b-a975-f0ce375ece6f. <https://directed-herding-research-platform.directed-herding-research.workers.dev>