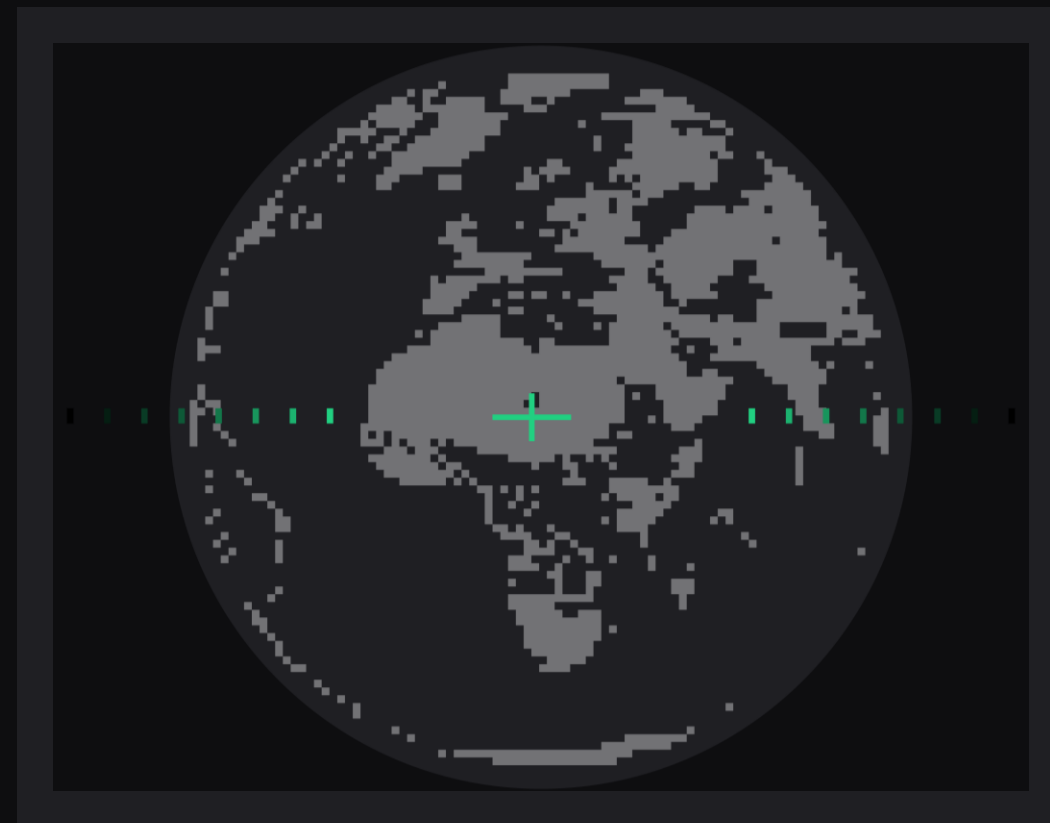


NOSIBLE WORLD V1.2

POINT-IN-TIME CORROBORATED EVENT INTELLIGENCE

A quant research appendix for coverage, entities, ontologies, seasonality, and reproducibility.

**15.3M**

CORROBORATED EVENTS

332.6M

SOURCE OBSERVATIONS

2015-2026

POINT-IN-TIME HISTORY

4,200

DAILY SLICES

HOW TO USE THIS APPENDIX

THE FIRST QUESTIONS A QUANT TEAM WILL ASK

01

What exactly is counted?

A point-in-time evidence contract that separates source observations from corroborated events.

02

How deep is the history?

4,200 daily slices from 2015 through 1 July 2026, with the partial year called out explicitly.

03

Where and in which languages?

Continental, country, and language distributions make geographic concentration visible.

04

Which companies and entities?

Ticker coverage, the NOSIBLE ontology, and named-entity breadth are shown together.

05

How is each event classified?

Fifteen ontology families expose editorial, sector, risk, cyber, disaster, and macro lenses.

06

Which temporal effects matter?

Weekend and year-end effects are measured before a researcher mistakes them for missingness.

READ THE EVIDENCE CONTRACT FIRST.

The charts that follow are descriptive, not sales claims. Definitions and reproducibility notes close the deck.

EVIDENCE CONTRACT

WE COUNT EVIDENCE, NOT HEADLINES

01 / OBSERVE

332.6M

Individual source observations retain the evidence behind each event.

02 / CORROBORATE

15.3M

Reports are resolved into point-in-time event records instead of counted as separate headlines.

03 / INDEX

15

Ontology families, entities, geography, language, and ticker mappings become researchable fields.

THE CONSEQUENCE

Every event-over-time chart in this appendix is specifically a count of events that could be corroborated point in time.

DATASET SCALE

15.3M EVENTS PRESERVE 332.6M SOURCE OBSERVATIONS

15.3M

CORROBORATED EVENTS

332.6M

SOURCE OBSERVATIONS

4,200

DAILY SLICES

21.7x

OBSERVATIONS PER EVENT

38.1K

UNIQUE SYMBOLS

76.6M

ENTITY ASSIGNMENTS

714.1M

ENTITY MENTIONS

15

ONTOLOGY FAMILIES

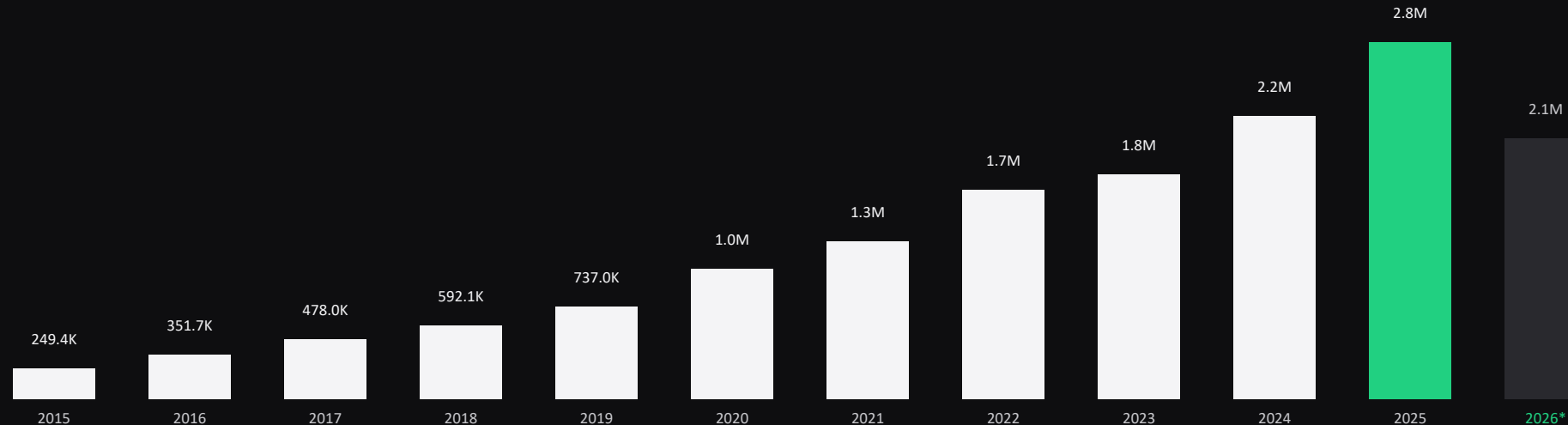
HISTORICAL DENSITY

THE HISTORY GETS DENSER AS CORROBORATION IMPROVES

POINT-IN-TIME HISTORY

2015-2026

Early history is intentionally sparser: only events satisfying the same corroboration contract are retained.



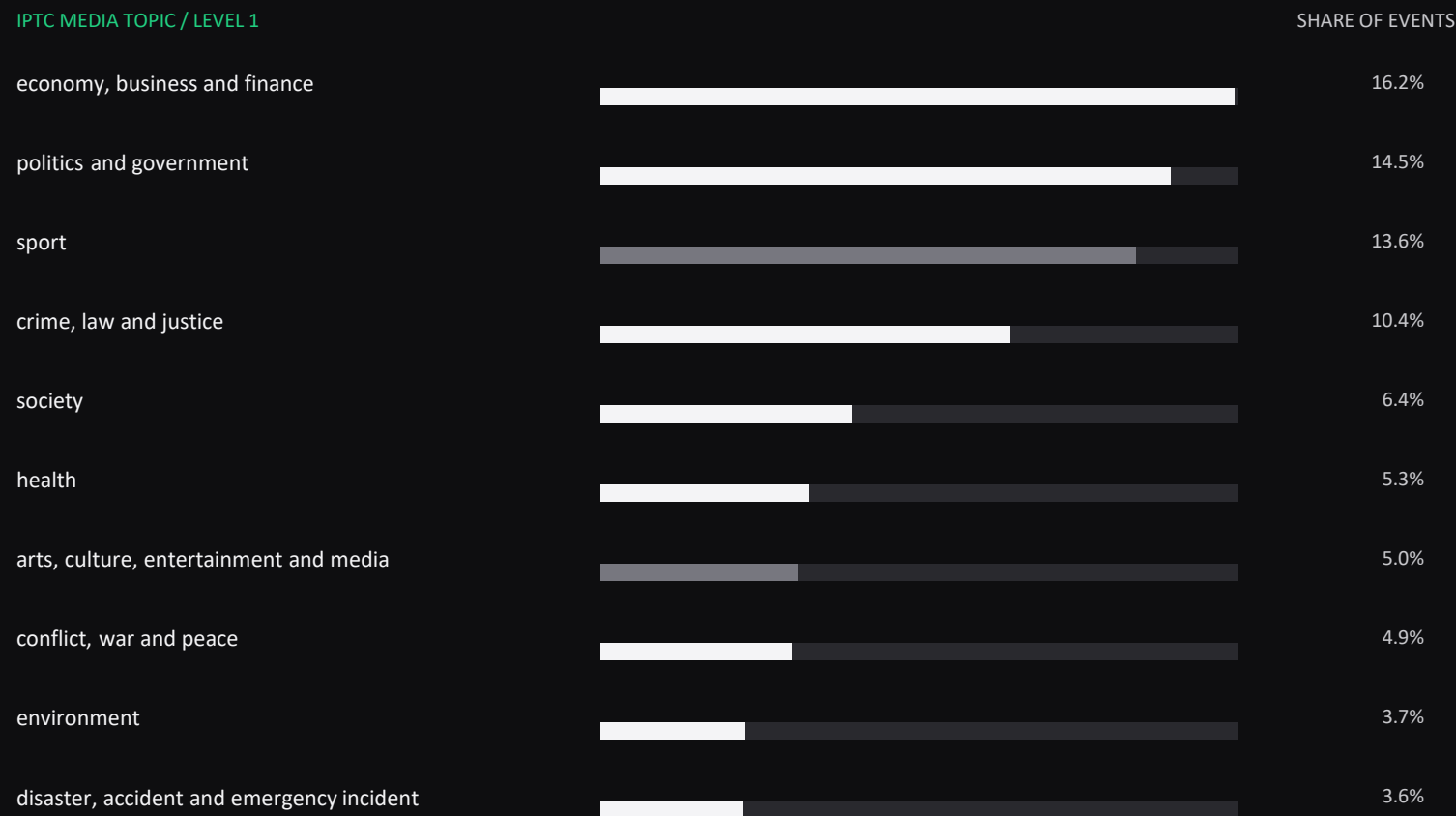
2.84M

Peak complete-year density in 2025. The 2026 bar is partial through 1 July and should not be annualized without an explicit model.

MATERIALITY

THE DATASET PRIORITIZES MATERIAL SIGNAL OVER MEDIA VOLUME

IPTC MEDIA TOPIC / LEVEL 1



46.0%

ECONOMY + POLITICS + CRIME + CONFLICT

DOWN-SAMPLING POLICY

Lower-signal IPTC media categories are downsampled where they are less likely to be material.

Examples include entertainment, culture, and sports. They remain represented for contextual and cross-sectional work.

Treat topic shares as a designed research universe, not an estimate of global media output.

GEOGRAPHIC BREADTH

75.7% OF EVENTS ORIGINATE OUTSIDE NORTH AMERICA

75.7%

OUTSIDE NORTH AMERICA

40.2%

ENGLISH-LANGUAGE

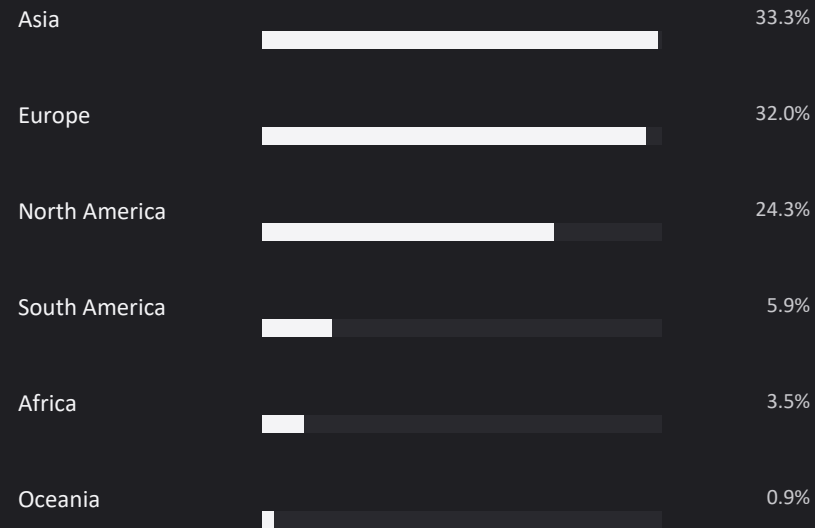
6

CONTINENTS

10

LEADING LANGUAGES SHOWN

CONTINENT SHARE



LEADING COUNTRIES

01	United States	21.7%
02	India	9.8%
03	Turkey	8.0%
04	Italy	4.6%
05	Brazil	3.7%
06	Spain	3.5%
07	United Kingdom	3.4%
08	Vietnam	3.3%

LEADING LANGUAGES

01	English	40.2%
02	Turkish	8.8%
03	Spanish	7.2%
04	Italian	5.0%
05	Portuguese	4.4%
06	Vietnamese	4.0%
07	Greek	3.8%
08	Russian	3.3%
09	Arabic	3.2%
10	German	2.8%

COMPANY COVERAGE

THE COMPANY LAYER IS BROAD, BUT COVERAGE IS DELIBERATELY SKEWED

2.2M

TICKERIZED EVENTS

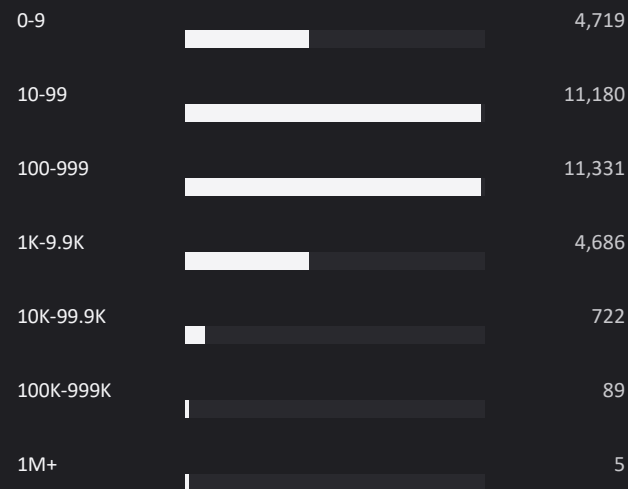
38.1K

UNIQUE SYMBOLS

14.3%

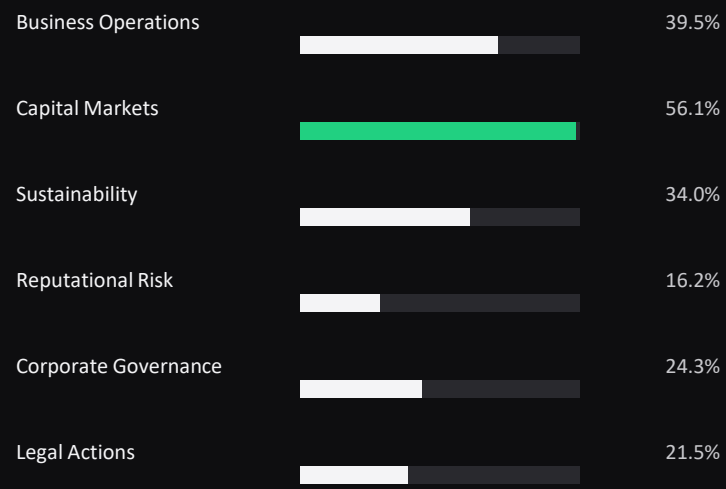
ALL EVENTS TICKERIZED

SYMBOLS BY SOURCE COVERAGE



A long tail is expected. Match thresholds to the universe and holding period under test.

NOSIBLE CATEGORY / TICKERIZATION RATE



Capital Markets and Investor Relations naturally carry the highest tickerization rates.

TOP SYMBOLS

		EVENTS / OBSERVATIONS
01	AAPL.US Apple Inc.	72K / 3.1M
02	DJT Trump Media & Technology Group ...	5K / 2.6M
03	6758.TSE Sony Corp	19K / 1.1M
04	TSLA Tesla Inc	21K / 1.1M
05	NYT New York Times Company	8K / 1.1M
06	AMZN Amazon.com Inc	23K / 951.7K
07	SPCX Space Exploration Technologies ...	13K / 874.7K

NAMED ENTITIES

THE ENTITY GRAPH IS MUCH LARGER THAN THE TICKER GRAPH

714.1M

TOTAL MENTIONS

76.6M

ENTITY ASSIGNMENTS

6.4M

UNIQUE PERSON STRINGS

3.2M

UNIQUE ORG STRINGS

NER TYPE	MENTIONS / UNIQUE STRINGS
PERSON	241.6M / 6.4M
ORG	195.0M / 3.2M
GPE	145.7M / 498.7K
LOC	64.8M / 972.4K
DATE	27.8M / 277.8K
PRODUCT	12.5M / 319.5K
MONEY	9.8M / 522.6K
EVENT	8.1M / 87.8K

Unique counts deduplicate case-sensitive canonical surface strings within each NER type.

TOP PERSON STRINGS

01	Trump	6.0M
02	Biden	1.3M
03	Donald Trump	810.3K
04	Putin	552.7K

TOP ORGANIZATION STRINGS

01	Apple	1.3M
02	NFL	1.1M
03	Google	960.9K
04	European Union	905.8K

ONTOLOGY COVERAGE

FIFTEEN ONTOLOGIES LET RESEARCHERS ASK DIFFERENT QUESTIONS

FAMILY / PRESCRIBED HIGHEST LEVEL / TOP THREE LABELS

CLASSIFICATION COVERAGE

EM-DAT uses Disaster subgroup because its root contains only two classes. Green cells are required WORLD fields.

SEASONAL EFFECTS

CALENDAR EFFECTS ARE LARGE ENOUGH TO MODEL EXPLICITLY

55.7%

WEEKEND DROP VS WEEKDAYS

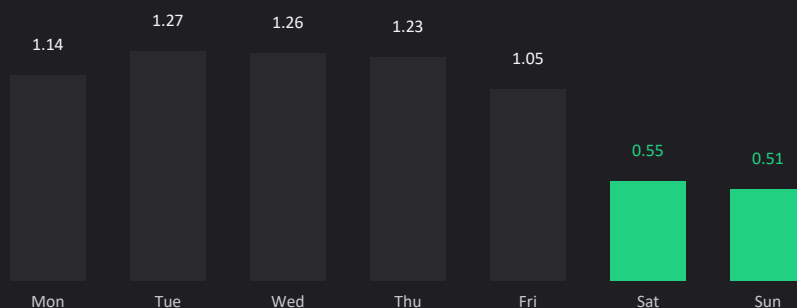
13.0%

20 DEC-10 JAN DROP / WEEKDAY ADJ.

2015-2025

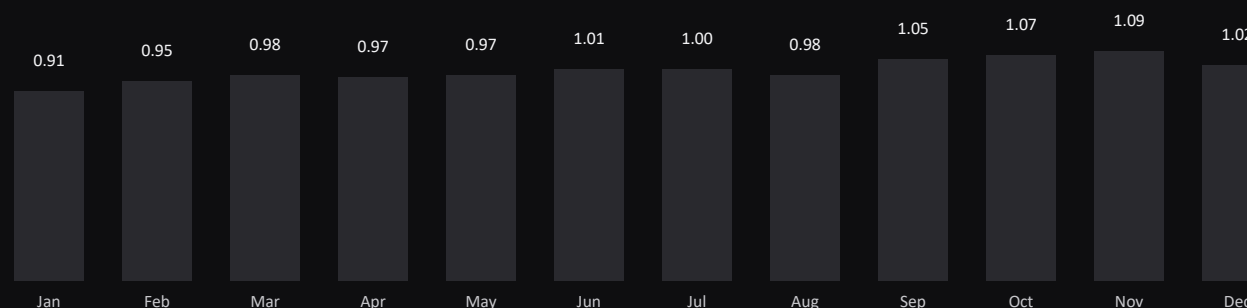
COMPLETE YEARS IN ESTIMATE

WEEKDAY INDEX / WITHIN-YEAR BASELINE = 1.00



Saturday and Sunday average 44.3% of weekday event density.

MONTH INDEX / WITHIN-YEAR BASELINE = 1.00



January is the weakest month; November is the strongest in the complete-year panel.

Research implication: include weekday and month controls before interpreting volume shocks. These effects are structural and should not be treated as random missingness.

RESEARCH AND REPRODUCIBILITY

THE RESEARCH STACK IS OPEN ENOUGH TO REPRODUCE

OUT-OF-SAMPLE STRATEGY

0.89 vs 0.64

Risk-on / risk-off Sharpe

2015-2026 untouched test; max drawdown improves from -34% to -18%.

PUBLISHED BENCHMARK

0.87 / 0.82

TPU levels / changes

Embedding-only reconstruction from 14.9M events and five sentences.

PUBLISHED BENCHMARK

0.90 / 0.79

Geopolitical risk

Daily levels / monthly changes against AI-GPR on 13.2M events.

FEATURE ENGINEERING

2 methods

Contrastive geometry

Training-free relevance and systemic-risk scores from frozen embeddings.

OPEN MODEL / FINANCIAL SENTIMENT

FINANCIAL SENTIMENT V1.1 BASE

Production-sized positive / neutral / negative company-impact classifier, with model and training dataset published on Hugging Face.

huggingface.co/NOSIBLE/financial-sentiment-v1.1-base

OPEN MODEL / FORWARD-LOOKING LANGUAGE

FORWARD-LOOKING V1.1 BASE

Separates expectations from realized outcomes, with model and training dataset published on Hugging Face.

huggingface.co/NOSIBLE/forward-looking-v1.1-base

READ THE METHODS

nosible.com/blog/turning-news-into-a-risk-on-risk-off-equity-signal nosible.com/blog/an-embedding-based-approach-to-trade-and-economic-policy-uncertainty nosible.com/blog/rebuilding-the-geopolitical-risk-index-from-nosible-world

QUANT DILIGENCE

A SERIOUS TRIAL SHOULD ANSWER THESE QUESTIONS FIRST

01

Point-in-time integrity

Rebuild features using only fields and references available at the decision timestamp.

02

Corroboration sensitivity

Compare event-level features with source-observation weighting and deduplication variants.

03

Taxonomy stability

Measure label drift, class prevalence, and null coverage through time and across languages.

04

Reference survivorship

Freeze ticker and website mappings by effective date before joining to a security master.

05

Calendar and missingness

Control for weekday, month, geography, language, and changing source density.

06

Leakage and look-ahead

Audit timestamps, event resolution windows, revisions, and any outcome-derived labels.

07

Capacity and latency

Measure event arrival, API/S3 throughput, feature freshness, and universe coverage.

08

Portfolio relevance

Run placebos, neutralization, turnover, costs, and true out-of-sample transfer.

MINIMUM VIABLE RESEARCH PROTOCOL

01

Freeze sample

02

Reproduce appendix

03

Define neutralization

04

Run placebos

05

Lock OOS test

DEFINITIONS, DISCLOSURES, AND REPRODUCIBILITY NOTES

TERM	DEFINITION	PRIMARY SOURCE
Event	Canonical point-in-time record created after resolving corroborating source observations.	events_map.ipc / one row per event
Source observation	One underlying report counted in total_coverage; retained as evidence for the event.	event.total_coverage
Tickerized event	Event carrying at least one symbol parsed from its ticker location offsets.	events.locs
Unique entity	Case-sensitive canonical surface string deduplicated within a single NER type.	canonical entities map
Ontology coverage	Share of all events assigned a value in the ontology family.	events_map.ipc facets
Seasonal index	Within-year normalized mean daily event count; 1.00 is that year's daily baseline.	complete years 2015-2025
Historical range	2015-01-01 through 2026-07-01; the 2026 annual bar is explicitly partial.	4,200 daily slices
Downsampling	Designed reduction of lower-materiality IPTC media categories while retaining representation.	materiality policy

AUDIT NOTES

NER top rankings exclude four audited non-entity tagging artifacts; aggregate totals remain unchanged. Entity and ticker parse failures in this run: zero.

REPRODUCIBILITY

All values are derived from committed daily slices. Ticker extraction reads location offsets rather than large embedding payloads. EM-DAT uses Disaster subgroup by design.