June 2, 2026

Binomial-Shannon-v2

News Sentiment Classification

Overview

What Shannon is

Binomial-Shannon-v2 (binomial-shannon-2) is an open-source NLP model that turns a piece of financial news — a wire story, a press release, a macro headline — into a structured information profile. Where a single-axis sentiment model returns one number, Shannon returns a full read: what kind of event this is, which direction it implies for the named ticker, how novel the information is, how specific the language is, how material it would be if true, and what kind of claim is being made.

It is the second model in Binomial Technologies' specialist zoo — a series of small, locally deployable ML models we publish, and that are also a part of our own trading models. The series consists of single-task specialists for quantitative finance: sentiment and event classification, market-regime classification, volatility forecasting, and other NLP/time-series tasks.
Each model in the zoo is named after a thinker who shaped how markets — or information itself — are understood. Shannon-2 is named after Claude Shannon, the founder of information theory. The name fits the job exactly: a news article is a noisy channel, and the question that matters is not "is this good or bad?" but "how much new, material, directional information does this message actually carry?" Shannon measures the signal in the wire.

The model is roughly 150M parameters — small enough to run on a CPU, fast enough to score the entire daily news flow on commodity hardware, and deterministic enough to be embedded directly inside a signal pipeline. It performs close to frontier reasoning models on the specific task of news characterization while running locally for free, cutting the API cost and latency of LLM-based news scoring by two orders of magnitude. A single shared encoder serves two modes through a learned router: a ticker mode that scores company-specific news against a named symbol, and a macro mode that classifies and scores market-wide news.

It is released under Apache 2.0 on HuggingFace and loads into any standard HuggingFace pipeline with trust_remote_code=True.

The numbers that make our case

- Implied-direction agreement vs. frontier panel: 0.877 Spearman — 98.8% of the 0.888 frontier-to-frontier ceiling
- Tone agreement vs. frontier panel: parity (0.905 vs. 0.883 ceiling)
- Directional sign accuracy vs. FinBERT (independent baseline): 95.8% vs. 62.0% on articles with a real directional signal
- Event-type agreement vs. frontier panel: 0.948 across the 10 binary event heads (~19 of 20)
- Macro mode: 0.814 18-way topic accuracy, +0.783 directional read, router accuracy ≈ 1.00
- Latency: ~15–30ms / call on CPU, low single-digit ms on GPU — deterministic, offline, zero-cost

Problem Solved

Financial news is one of the highest-frequency information events in markets. Tens of thousands of articles, filings, and press releases hit the wire every trading day. The structured facts in a headline — an EPS number, a guidance figure, a deal price — are priced in within minutes. The unstructured language layer is where the slower, exploitable signal lives:
"the company announced a strategic review of its portfolio" vs. "the company confirmed it has retained advisors to explore a sale" vs. "people familiar with the matter say a sale is unlikely in the near term"
These three sentences carry very different information about the same situation — different event types, different directions, different novelty, and very different claim status (announced fact vs. confirmed fact vs. unconfirmed rumor). A systematic desk that can separate them at scale has an edge. It is also where the available approaches have historically struggled, because each has a structural weakness.

Dictionary and lexicon methods

The traditional approach: dictionary methods such as Loughran-McDonald, bag-of-words classifiers, simple keyword rules. Cheap, fast, and explainable. What it misses:

- Context — "missed estimates" is a hard negative for a company that guided up last quarter and an expected non-event for one that pre-announced. A dictionary scores them identically.
- Nuance — a layoff announced in upbeat corporate prose ("optimizing our cost structure to drive long-term value") is tonally positive and directionally negative. A lexicon reads the words, not the implication.
- Schema — a dictionary returns a scalar. It cannot return "this is an M&A rumor, low novelty, high materiality if true."

Single-axis sentiment models (FinBERT and kin)

A real step up: a transformer fine-tuned for financial sentiment. But it still collapses a rich event into a single positive / neutral / negative axis, and on real wire news it abstains to neutral roughly half the time — the least useful answer on an article that actually carries directional information. It has no event taxonomy, no novelty estimate, no claim-type distinction, and no separation of how-it's-written (tone) from what-it-implies (direction). A press release announcing a buyback and one disclosing a lawsuit can both come back "neutral."

Frontier LLMs over an API

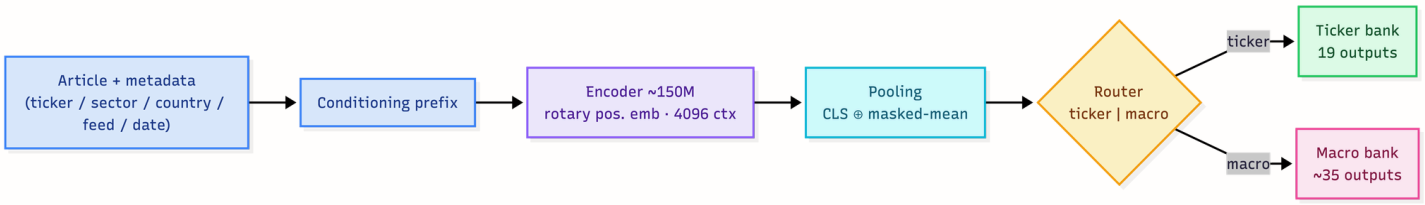
The new default: prompt Claude, GPT, Grok, or Gemini with a structured-output schema and let the model characterize each article. Quality is strong, but there are real constraints:

- Cost — even at low reasoning effort, frontier APIs run roughly $0.002–$0.02 per article. Across a full daily news flow this compounds fast: a desk scoring 50,000 articles a day is at four-to-five figures a week before retries.
- Latency — a reasoning-capable model takes seconds per article. Scoring the news flow in near-real-time requires heavy parallelization and capacity management.
- Schema instability — outputs drift across model updates; a float field today can become a string label or fail validation tomorrow, forcing retry layers, schema guards, and version pins just to stay consistent.

Architecture

Shannon-2 is a ~150M-parameter transformer encoder with a learned mode router and two banks of lightweight regression / classification heads sharing a pooled representation.

The pipeline has five stages: input conditioning → encoder → pooling → router → the routed head bank.



Input conditioning

The model takes an article with a small amount of structured metadata and renders it as a deterministic prefix prepended to the text. In ticker mode:

[SOURCE: news] [SECTOR: Technology] [COUNTRY: US] [TICKER: NVDA] [DATE: 2026-01-29]
TITLE: ... BODY: ...

In macro mode the prefix carries the feed and source instead of a ticker:
[SOURCE: MACRO] [FEED: central-bank] [SITE: reuters] [DATE: 2026-01-29]
TITLE: ... BODY: ...

This is not a label appended to the input — the model is trained on prefixed inputs from the start, so the encoder learns to weight language differently given the named ticker, its sector, and its country. The same sentence about a regulatory probe produces a sharper negative read for a name whose franchise depends on that regulator, and an attenuated one where it is peripheral. The ticker tag also scopes the read: an article that mentions five companies is scored only against the named symbol, so the same wire story can yield five different directional reads depending on which ticker it is conditioned on.

Why conditioning matters

Conditioning is what lets one model serve company news and market news from a single set of weights. Without the ticker scope, a multi-company article has no defined target and the directional read is undefined. Without sector/country, context-sensitive judgments — is this magnitude meaningful or noise for this kind of business? — collapse toward a generic prior.

Operational implication

Conditioning inputs (source, sector, country, ticker) are not optional in practice. Defaults let inference run, but degrade signal quality and, in ticker mode, leave the directional read unscoped. In production this implies a dependency on a clean security-master and feed lookup before scoring.

Encoder

The encoder is a transformer-based model with rotary positional embeddings, initialized from a strong public checkpoint and adapted for this task. It is trained with a 4,096-token context, though inference typically runs at ~1,024 tokens because news bodies are short — the full context is rarely needed, which is part of why the model is so fast. It produces a hidden state for every token; the head banks and router together account for a small fraction of the total parameter count, with the bulk residing in the encoder.

Pooling

After the encoder, the token-level hidden states are collapsed into a single fixed-size representation with two pooling mechanisms:

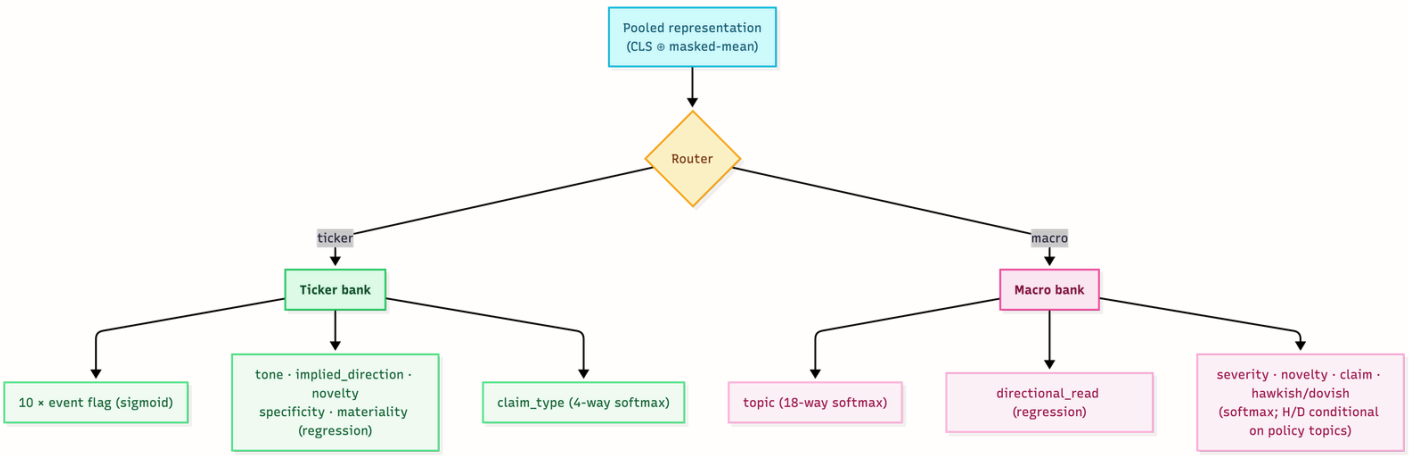
- CLS pool — the hidden state of the summary token at position 0; the model's learned global summary of the article.
- Masked-mean pool — the average of all non-padding token hidden states; the distributed signal across the whole article.

The two vectors are concatenated and passed to the router and heads. Short headlines concentrate signal in the summary token; longer stories benefit from aggregation across tokens. Using only one pooling method degrades one end of the length spectrum.

Router and head banks

A 2-way classifier on the pooled representation decides whether the article is company-specific (ticker mode) or market-wide (macro mode) and dispatches it to the matching head bank. The router is near-perfect in evaluation (accuracy ≈ 1.00), which matters operationally: a caller can hand Shannon any article without pre-classifying it, and the model self-selects the correct output schema. Loss is masked per mode during training, so each head bank only ever learns from articles of its own type.
Each head is a lightweight 2-layer MLP (Linear → GELU → Dropout → Linear) over the shared pooled representation:

- Ticker bank — 19 outputs: 10 binary event flags (sigmoid + BCE), five regression scores (tone, implied_direction in [-1, +1], novelty, specificity, materiality_if_true in [0, 1]), and a 4-way claim_type softmax.
- Macro bank — ~35 outputs: an 18-way topic softmax, a signed directional_read in [-1, +1], a 5-way severity scale, a 3-way macro-novelty class, a 4-way macro-claim class, and a 5-way hawkish_dovish class that is conditional on a monetary-policy / rates topic (loss-masked otherwise).



The training data

The model is trained on hundreds of thousands of ticker-tagged news articles and press releases spanning a broad global equity universe, dated 2018 through 2026, plus a large corpus of market-wide / macro news for the macro mode. Each ticker article carries sector and country metadata; each macro article carries feed and source metadata.

Coverage by source

- Wire / general financial news: Majority of the ticker corpus
- Company press releases: Substantial minority, scored under the same schema
- Macro / market-wide news: Separate corpus routed to the macro mode

Coverage by region

Heavily US-tilted, with meaningful developed-market coverage (UK, Western Europe, developed Asia-Pacific) and lighter emerging-market representation. The ticker universe skews toward the analyst-followed, liquid names that dominate news flow.

Coverage by time

- 2018–2021: Broad; pre- and early-COVID news regime
- 2022–2024: Full; rate cycle, macro-headline-heavy
- 2025–2026: Full through the training window; most recent months held out

Labels

Labels are generated by a frontier reasoning system applied uniformly across the full training corpus under a fixed scoring rubric, then validated two ways on a held-out benchmark: against an independent panel of frontier LLMs (none of which is the teacher) and against FinBERT, an independent public baseline the model never learned from. No human annotation is used. The model is trained as a language-imitation system — learning to replicate structured outputs produced by stronger, slower models — and is then checked against judges and a baseline it did not learn from, so the evaluation is not merely self-referential.

Split

- Method: Forward-temporal
- Train: Articles dated on or before 2025-09-30
- Test: Articles dated 2025-10 through 2026-05
- Unit: Article

Although the labels are deterministic functions of the text (produced by a language model, not future market data), Shannon uses a forward-temporal split rather than a random one. This is deliberately stricter than necessary for leakage: it tests whether the model generalizes to genuinely future news — new companies, new events, new language — rather than to a random holdout drawn from the same period it trained on. Every evaluation number in this document is measured on articles dated strictly after the training window.

What it returns

Given an article with metadata, Shannon-2 returns one of two structured profiles depending on the routed mode.

Ticker mode — 19 fields

- Event flags (10 binary, multi-label): earnings, guidance, m_and_a, regulatory_legal, product, exec_change, dividend_buyback, analyst_rating, macro_sector, other
- tone [-1, +1]: How the article is written — hostile to laudatory — independent of whether the news is good or bad
- implied_direction [-1, +1]: What the article implies for the named ticker if its claims hold — the headline directional signal, deliberately separate from tone
- novelty [0, 1]: New first-reporting vs. a rehash of the prior day's record
- specificity [0, 1]: Vague hand-waving vs. fully quantified prose
- materiality_if_true [0, 1]: How much it would move the thesis for this company if the claims hold
- claim_type (4-way softmax): fact / opinion / rumor / forecast

Macro mode — ~35 fields

- topic (18-way): monetary_policy, inflation, growth, labor, rates_fixed_income, equities_markets, fx_currency, energy, commodities, credit_banking, crypto, trade_policy, geopolitics, …
- directional_read [-1, +1]: Signed read for risk assets
- Categorical: severity (5-way), novelty_macro (3-way), claim_macro (4-way), hawkish_dovish (5-way, conditional on policy topics)

The shape is JSON-serializable, deterministic, and stable. Versions within v2 will not change the schema; a later version may add dimensions but will preserve the layout for backward compatibility.

Worked examples

Macro mode.

A central-bank wire story — "Policymakers signaled two further cuts this year as inflation cooled toward target" — auto-routes to macro mode.

A single-axis sentiment model reads the same story as roughly neutral — there is no overtly positive or negative language — and misses the directional read (+0.29, net supportive of risk assets), the policy lean (mildly_dovish), and the novelty entirely.


```
{
  "_meta": {
    "model": "binomial-shannon-2",
    "mode": "macro",
    "feed": "central-bank",
    "site": "reuters",
    "date": "2026-01-29"
  },
  "topic": { "monetary_policy": 0.93, "rates_fixed_income": 0.04, "inflation": 0.02 },
  "directional_read": 0.29,
  "severity":        { "noise": 0.02, "minor": 0.11, "notable": 0.71, "major": 0.15, "crisis": 0.01 },
  "novelty_macro":   { "rehash": 0.08, "commentary": 0.17, "breaking": 0.75 },
  "claim_macro":     { "fact": 0.88, "opinion": 0.06, "rumor": 0.02, "forecast": 0.04 },
  "hawkish_dovish": { "dovish": 0.18, "mildly_dovish": 0.62, "neutral": 0.16, "mildly_hawkish": 0.03, "hawkish": 0.01 }
}
```

Ticker mode:

A company press release — "[TICKER: XYZ] announces restructuring; to reduce workforce by 8% to streamline operations" — is auto-routed to ticker mode and returns the full 19-field profile.


```
{
  "_meta": {
    "model": "binomial-shannon-2",
    "mode": "ticker",
    "ticker": "XYZ",
    "sector": "Technology",
    "country": "US",
    "date": "2026-01-29",
    "source": "press_release"
  },
  "events": {
    "earnings":          { "mentioned": false, "prob": 0.04 },
    "guidance":          { "mentioned": false, "prob": 0.07 },
    "m_and_a":           { "mentioned": false, "prob": 0.02 },
    "regulatory_legal": { "mentioned": false, "prob": 0.03 },
    "product":           { "mentioned": false, "prob": 0.06 },
    "exec_change":       { "mentioned": false, "prob": 0.09 },
    "dividend_buyback": { "mentioned": false, "prob": 0.02 },
    "analyst_rating":   { "mentioned": false, "prob": 0.01 },
    "macro_sector":      { "mentioned": false, "prob": 0.12 },
    "other":             { "mentioned": true,  "prob": 0.91 }
  },
  "tone": 0.10,
  "implied_direction": -0.35,
  "novelty": 0.78,
  "specificity": 0.70,
  "materiality_if_true": 0.40,
  "claim_type": { "fact": 0.86, "opinion": 0.05, "rumor": 0.03, "forecast": 0.06 }
}
```


This is the case a lexicon or single-axis model gets exactly wrong: the words are positive (tone +0.10), the implication is negative (implied_direction -0.35), and Shannon separates the two — while also recording that this is a high-novelty, quantified (specificity 0.70), factual disclosure.

Evaluation

Methodology

Shannon-2's ticker heads are evaluated against a panel of frontier reasoning systems on a frozen, held-out benchmark of 500 articles drawn from the forward-temporal test window (dated strictly after every training example)

- Grok-4.2 reasoning (xAI)
- Claude Opus 4.7 (Anthropic)
- Shannon-2 (Binomial AI Research)

The same 500 articles are scored by all systems under Shannon's exact rubric. We then measure how well Shannon agrees with the panel relative to how well the panel members agree with each other. Two frontier models, handed the same article and rubric, do not agree perfectly — subjective axes admit genuine judgment spread — so the rate at which they agree with each other is the ceiling: the highest score any model could realistically post. Neither panel model is Shannon's teacher, so this is an external check, not a measure of how faithfully the student imitates one specific instructor. Alongside the panel, the full holdout is scored for per-head self-evaluation, and an independent FinBERT baseline confirms the directional quality is real and not memorized.

Why frontier–frontier agreement isn't 1.0

If we use frontier LLMs as graders, why don't they agree perfectly with each other?

1. Genuine ambiguity. Whether an article is "breaking" or a "rehash" of yesterday's record is genuinely contestable — which is why novelty has the lowest ceiling of any axis.
2. Calibration differences. Different models anchor the "neutral" point of a directional read differently. This is calibration drift, not noise.
3. Different reasoning styles. Models trained with different objectives reach different conclusions on edge cases — particularly on materiality, where the answer depends on how much company context the rater brings.

These factors cap achievable agreement at roughly 0.89 on implied direction and lower on subjective axes like novelty (≈0.68), even between equally capable frontier systems.

The headline result

Shannon-2 reproduces 98.8% of the agreement that frontier reasoning systems have with each other on directional news scoring, and sits at parity on tone. For a desk that scores the full news flow, this is the difference between a five-figure quarterly API line and a rounding error.

Implied direction — panel agreement (Spearman, 500-article benchmark):

Comparison	Spearman
Frontier ↔ Frontier ceiling (Grok-4.1 ↔ Opus 4.7)	0.888
Shannon-2 ↔ Frontier panel mean	0.877
% of ceiling	98.80%
Shannon-2 ↔ Grok-4.2	0.859
Shannon-2 ↔ Opus 4.7	0.847

Shannon agrees with the panel almost exactly as well as the two frontiers agree with each other. The two per-frontier rows are close, which matters: a small model that had simply overfit one frontier's idiom would track that model and diverge from the other. Shannon tracks both about equally — the signature of a model that learned the underlying judgment rather than one instructor's surface style.

The full pairwise matrix makes the structure explicit. Shannon-2 is included as a third voter on equal footing, so the whole matrix is interpretable on the same scale:

↓ vs →	Grok-4.2	Opus 4.7	Shannon-2
Grok-4.2	—	0.888	0.859
Opus 4.7	0.888	—	0.847
Shannon-2	0.859	0.847	—

Shannon-2's worst pair (0.847) is barely below the frontier ceiling (0.888): the gap between "how well Shannon agrees with a frontier" and "how well two frontiers agree with each other" is about 0.04 in rank correlation — small enough that, on a blind read of the matrix, Shannon does not stand out as the non-frontier row.

Tone — parity with frontier (Spearman):

Comparison	Spearman
Frontier ↔️ Frontier ceiling	0.883
Shannon-2 ↔️ Frontier panel mean	0.905
Shannon-2 ↔️ Grok-4.2	0.884
Shannon-2 ↔️ Opus 4.7	0.869

On tone Shannon lands at — and slightly past — the frontier-to-frontier rate. The >100% is not magic: comparing against the panel mean cancels some of the idiosyncratic noise each individual frontier carries, so a well-calibrated small model can track the consensus as tightly as either frontier tracks the other. The honest reading is frontier-parity, not super-human.

Event mention and claim type

The ten binary "what kind of event is this?" heads are easier than the regression heads, and the model performs proportionally better. On the panel benchmark, mean event-mention agreement is 0.948 — the model agrees with the panel on event type about 19 times out of 20. On the full holdout the event heads reach a macro-F1 of 0.791, and claim-type classification is 89.5% accurate across the four-way fact / opinion / rumor / forecast distinction.

Per-event F1 (full holdout):

Event	Mention F1	Base Rate
regulatory_legal	0.92	8.50%
earnings	0.87	15.60%
guidance	0.84	3.70%
m_and_a	0.81	3.80%
exec_change	0.77	1.70%
other	0.77	52.50%
dividend_buyback	0.75	3.70%
product	0.75	11.60%
analyst_rating	0.72	10.10%
macro_sector	0.72	20.30%

Per-head regression (full holdout, Pearson vs. teacher):

Head	Pearson
implied_direction	**0.854**
tone	0.834
materiality_if_true	0.788
specificity	0.778
novelty	0.62

Novelty is the weakest axis — consistent with the panel, where novelty also has the lowest frontier-to-frontier ceiling. Deciding whether an article is first-reporting or a rehash is genuinely contestable, and every model's number on that axis, Shannon's included, should be read as soft. Lining every continuous axis up against its own frontier ceiling shows where Shannon sits relative to the achievable maximum (Spearman, on the 500-article panel):

Axis	Frontier Ceiling	Shannon vs Panel	Δ vs Ceiling
implied_direction	0.888	0.877	-0.011
tone	0.883	0.905	0.022
specificity	0.829	0.882	0.053
materiality_if_true	0.831	0.794	-0.037
novelty	0.684	0.719	0.035

On four of five axes Shannon is within ±0.04 of the ceiling or above it; materiality is the only axis where it trails by a meaningful margin, and even there it lands at 96% of the frontier-to-frontier rate.

Scorecard — the headline metrics in one place:

Metric	Shannon-2
Implied-direction, % of frontier ceiling (panel)	98.80%
Tone, vs frontier ceiling (panel)	Parity
Event-mention agreement (panel)	0.948
Event macro-F1 (holdout)	0.791
Claim-type accuracy (holdout)	0.895
Directional sign accuracy vs FinBERT	95.8% vs 62.0%
Macro topic accuracy (18-way)	0.814
Macro directional-read Pearson	0.783
Router accuracy	≈ 1.00

The independent baseline: Shannon vs. FinBERT

This is the metric that does not depend on the teacher at all. On the subset of holdout articles where the panel's directional magnitude is at least 0.2 — articles with a genuine directional signal — we compare Shannon and FinBERT, an independent public sentiment model trained on different data with a different objective:

Metric	Shannon	FinBERT
Directional sign accuracy	95.80%	62.00%
Pearson vs. panel implied_direction	0.862	0.562
N (amount)	5,615	5,615

Shannon agrees with the panel on direction 96% of the time; FinBERT, 62%. The gap is structural: FinBERT is single-axis and abstains to neutral on roughly half of real wire news, which on a directional article is a disagreement. Shannon roughly halves FinBERT's directional error rate while returning eighteen additional fields it does not produce at all.

Macro mode — topic and direction

Macro mode is evaluated on the full macro holdout. The headline is 18-way topic accuracy of 0.814 and a directional_read correlating +0.783 (Pearson) with the panel. Per-topic F1 is strongest on the well-defined market topics and weakest on the diffuse long tail:

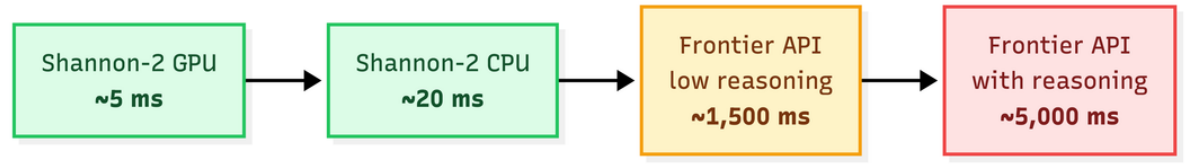
Topic	F1	Support
commodities	0.94	2,061
equities_markets	0.88	3,613
fx_currency	0.88	3,861
trade_policy	0.81	400
monetary_policy	0.79	1,662
inflation	0.7	526
growth	0.66	621
geopolitics	0.48	346
technicals	0.25	517

Technicals and geopolitics are the weakest: both are catch-all categories where even frontier labels are inconsistent. The well-posed market topics — commodities, equities, FX, monetary policy — are where the macro mode is reliable enough to gate on.

Costs and latency

Approach	Latency / Article	Cost / Article	Determinism	Offline?
Frontier LLM API (low reasoning)	1–5 s	$0.002–$0.02	No (sampling)	No
Shannon-2 on a modern CPU	~15–30 ms	~$0 (electricity)	Yes	Yes
Shannon-2 on a single GPU (bf16)	Low single-digit ms	~$0	Yes	Yes

The latency gap is two to three orders of magnitude — the difference between a real-time news block and a batch job that has to be parallelized and rate-managed:



Total cost of ownership

For a desk scoring its news flow, the cost ratio is roughly two orders of magnitude:

News Flow	Articles / Day	Frontier API (~$0.008/article)	Shannon-2
Single-name watchlist	1,000	~$8/day	~$0
Sector / desk universe	20,000	~$160/day	~$0
Full market news flow	100,000	~$800/day (~$200k/yr)	~$0
Historical backfill (multi-year)	50M+	$400k+	~$0

These assume one scoring pass per article. A realistic workflow that re-scores the archive as the schema evolves multiplies the API line accordingly while leaving the Shannon line essentially flat. Determinism — the same article always produces the same profile — makes downstream backtesting reproducible without freezing API responses or managing version pins.

Time-to-signal in a live pipeline

Approach	Wall-clock for 5,000 Articles in a Burst
Frontier LLM API (parallelized, 16-way)	Several minutes plus retry overhead
Shannon-2 on a single CPU instance	A couple of minutes
Shannon-2 on a single GPU instance	Seconds

For a desk that needs the news block scored quickly to inform positioning, Shannon completes essentially in real time relative to news arrival, with no rate limits or capacity management.

Putting it to work

Shannon-2 outputs are features, not trades. They are designed to be consumed by existing quant infrastructure.

Scoring an article. Shannon-2 is a regular transformers model — it loads with trust_remote_code=True and exposes a predict() method that runs the router and both head banks. The caller prepends the conditioning prefix; the router selects the schema:

```
import torch
from transformers import AutoModel, AutoTokenizer

REPO = "BinomialTechnologies/binomial-shannon-2"
tok = AutoTokenizer.from_pretrained(REPO)
model = AutoModel.from_pretrained(REPO, trust_remote_code=True).eval()

prefix = "[SOURCE: news] [SECTOR: Technology] [COUNTRY: US] [TICKER: NVDA] [DATE: 2026-01-29]\n\n"
text = prefix + f"TITLE: {title}\n\nBODY: {body}"
enc = tok(text, truncation=True, max_length=1024, return_tensors="pt")

with torch.no_grad():
    out = model.predict(**enc)              # dict of tensors

out["mode_prob"].argmax(-1).item()          # 0 = ticker, 1 = macro
float(out["implied_direction"])             # e.g. -0.35
float(out["materiality_if_true"])           # e.g.  0.40
```

Batched scoring. For a day's news flow, build the prefixes per row and score in batches — a single GPU clears tens of thousands of articles a minute:

```
texts = [build_prefix(r) + f"TITLE: {r.title}\n\nBODY: {r.body}" for r in rows]
enc = tok(texts, truncation=True, max_length=1024, padding=True, return_tensors="pt")
with torch.no_grad():
    out = model.predict(**enc)
df["implied_direction"]    = out["implied_direction"].tolist()
df["novelty"]              = out["novelty"].tolist()
df["materiality_if_true"] = out["materiality_if_true"].tolist()
```

Deployment patterns

Beyond a notebook, the same model drops into any standard HF surface: a batch job over the news archive, a FastAPI microservice wrapping the scorer, HuggingFace Inference Endpoints, or a subprocess in a data pipeline for environments that cannot import torch directly. The router means a caller hands the model any article and gets the correct schema back automatically.

- Real-time news triage — score the inbound flow and gate on implied_direction, materiality_if_true, and novelty to surface the small set of articles that carry new, material, directional information, separating first-reporting from rehash with no return-data dependency.
- Event-study triggers — use the event flags and directional read as clean event boundaries: all articles where m_and_a fired and implied_direction > 0.5 (explicit positive deal news), or regulatory_legal with a negative read (adverse legal events).
- News-flow sentiment indices — aggregate ticker-mode directional reads across a name, sector, or index into a deterministic, backfillable real-time sentiment series.
- Macro nowcasting — route the macro feed through macro mode to build topic-tagged, direction-scored, severity-weighted series for monetary policy, inflation, growth, and trade — a structured macro-news overlay traditional indices do not provide.
- Claim and novelty filtering — use claim_type and novelty to down-weight rumor and rehash before a signal hits the book, treating an unconfirmed rumor differently from a confirmed fact, which a single-axis model cannot distinguish at all.

Screening filter

Because the outputs are deterministic and on a fixed scale, a desk can pre-filter the flow with a plain SQL predicate over a materialized signals table:

```sql
SELECT ticker, date FROM shannon_signals
WHERE  implied_direction    >  0.3     -- positive read
  AND  materiality_if_true  >  0.5     -- moves the thesis
  AND  novelty              >  0.6     -- genuinely new
  AND  claim_type           =  'fact'; -- not rumor / opinion
```

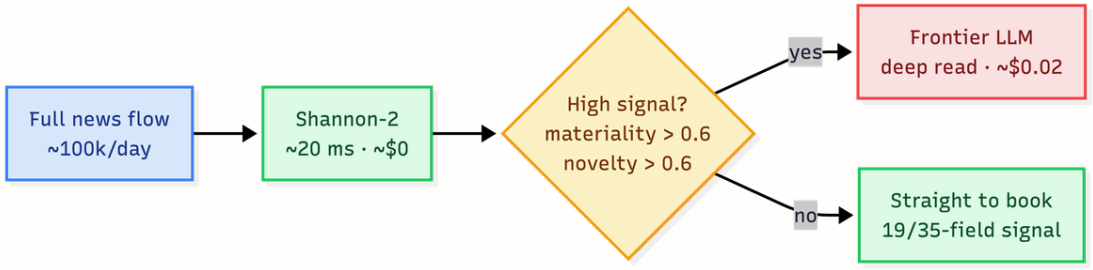
Aggregation into an index

Group ticker-mode reads into a sector or index series that backfills across the full archive and updates as news lands:

```python
sector_sentiment = (
    df.group_by(["sector", "date"])
      .agg(pl.col("implied_direction").mean().alias("net_read"))
)
```

Triage routing

A common pattern uses Shannon as a cheap first pass over the entire flow, escalating only the high-signal minority to an expensive frontier LLM for a deep read:



Limitations

- Macro long tail — diffuse catch-all topics (technicals F1 0.25, geopolitics 0.48) are far weaker than the well-posed market topics. Gate on the strong topics; treat the long tail as advisory.
- English only — training data is heavily English and US-tilted. Non-English news degrades, especially in automated translation.
- Short context at inference — ~1,024 tokens covers nearly all wire news and press releases, but long-form research notes exceed it and lose middle content to truncation.
- Distillation framing — the model is a language-imitation system and inherits both strengths and biases from its teacher. The panel and FinBERT comparisons measure differences between models, not shared miscalibration.
- Characterizer — Shannon reports what an article says and implies; it does not verify the claim or predict the realized return. A claim_type of fact means the article presents the claim as fact, not that Shannon has confirmed it. It is a feature extractor, not a standalone trading system, and ships as a Tier 2 research preview rather than a return-validated product.

Where to find it

- Model weights & card: huggingface.co/BinomialTechnologies/binomial-shannon-2
- License: Apache 2.0