DISSENSUS WORKING PAPER SERIES
DAI-2508

Are Whitepaper Claims Reflected in Market Structure?

A Contamination-Aware Pipeline and a Power-Limited Null

Murad Farzulla[1,2,*]

[1]Dissensus, London, UK [2]King's College London, London, UK

[*]Correspondence: murad@dissensus.ai ORCID: 0009-0002-7164-8704

June 2026

Abstract

Do the functional narratives in cryptocurrency whitepapers correspond to how their tokens behave in markets? We develop a content-verified, contamination-aware pipeline for measuring structural correspondence between project narratives and market structure, and report two results. The first is a cautionary one. An apparent entity-level signal in an earlier version of our corpus—specialised tokens appearing to align more strongly than broad infrastructure tokens—was entirely an artefact of corpus contamination: roughly a quarter of the documents were failed-download stubs or wrong-document whitepapers (for example, a "Cosmos" entry that was in fact Binance Smart Chain text), and the apparent ordering does not survive content verification: on the clean corpus no token registers as helping alignment. We therefore report it as a contamination diagnosis, not a finding. The second is an honest null. Combining zero-shot NLP classification of 43 content-verified whitepapers across 10 semantic categories with seven cross-sectional market-structure statistics computed from hourly data (17,543 timestamps, 2023–2024), and aligning the two spaces with Procrustes rotation and Tucker's congruence coefficient (ϕ), we do not detect a significant claims–market alignment in this $n = 43$ sample (dimension-matched $\phi = 0.303$, our primary estimate; zero-padded $\phi = 0.223$, a conservative display floor; both non-significant). A positive-control and power analysis shows the binding constraint is the low reliability of the text instrument: at the estimator-relevant within-category reliability the minimum detectable effect is $\phi \approx 0.44$, still well above the matched-dimension estimate of 0.303. This is absence of evidence for alignment, not evidence of its absence: our results are inconsistent with strong alignment ($\phi \geq 0.70$), which the design has over 80% power to detect, but cannot distinguish weak alignment ($\phi \approx 0.3$) from none. A positive-control simulation confirms the estimator recovers injected cross-domain congruence almost exactly, so the non-detection reflects limited power rather than an insensitive pipeline. We frame the contribution as a method plus a cautionary tale for text-based studies of narrative–market structural correspondence, which routinely operate below an unstated detectability floor.

Keywords: Cryptocurrency, Narrative Economics, NLP, Procrustes Rotation, Tucker's Congruence Coefficient, Zero-Shot Classification, Power Analysis, Corpus Contamination

JEL Codes: G14, G12, C38, C45

Acknowledgements. The author acknowledges Claude (Anthropic) for assistance with pipeline development, mathematical exposition, and technical writing. All errors, omissions, and interpretive limitations remain the author's responsibility.

Data & Code Availability. Reproducible code and data are available at `https://github.com/studiofarzulla/whitepaper-claims`.

1 Introduction

Cryptocurrency projects typically articulate their value propositions at inception, in whitepapers that make explicit claims about functionality, use cases, and technical architecture. Unlike equities, whose

fundamentals emerge gradually through earnings reports and analyst coverage, these founding documents are often the first and most detailed statement of what a token is *for*, and they should, in principle, relate to how the asset subsequently behaves in markets. Whether they do is an open question. The efficient-market view (Fama, 1970) holds that informative narratives are quickly impounded into prices; Shiller (2017) argues instead that "narrative economics" can decouple prices from fundamentals; and Aste (2019) documents significant price–sentiment correlations across nearly two thousand cryptocurrencies.

This tension motivates our research question: *are whitepaper claims reflected in market structure?* We measure contemporaneous structural correspondence between two representational spaces—a claims matrix $\mathbf{C} \in \mathbb{R}^{N \times K}$ obtained from zero-shot classification of whitepaper text across $K = 10$ semantic categories, and a market-statistics matrix $\mathbf{S} \in \mathbb{R}^{M \times J}$ of $J = 7$ cross-sectional financial metrics—aligning them with Procrustes rotation and comparing them with Tucker's congruence coefficient ϕ. This is a test of structural alignment between two spaces, not of prediction or forecasting.

We report two results. The first is a *cautionary tale*. An apparent entity-level signal in an earlier version of our corpus—specialised tokens appearing to align more strongly than broad infrastructure tokens—was entirely an artefact of corpus contamination. Roughly a quarter of the documents were failed-download stubs or wrong-document whitepapers (for example, a "Cosmos" entry that was in fact Binance Smart Chain text), and the apparent ordering does not survive content verification: on the clean corpus no token registers as helping alignment. We report it as a contamination diagnosis, not a finding. The second is an *honest null*. On $n = 43$ content-verified whitepapers matched to hourly market data (17,543 timestamps, 2023–2024), we do not detect a significant claims–market alignment (dimension-matched $\phi = 0.303$, zero-padded $\phi = 0.223$; both non-significant). A positive-control and power analysis shows the binding constraint is the low reliability of the text instrument: at the estimator-relevant within-category reliability the realistic minimum detectable effect is $\phi \approx 0.44$, still well above the matched-dimension estimate of 0.303. This is absence of evidence for alignment, not evidence of its absence: our results are inconsistent with strong alignment ($\phi \geq 0.70$), which the design has over 80% power to detect, but cannot distinguish weak alignment ($\phi \approx 0.3$) from none.

Contributions. (1) We introduce a content-verified, contamination-aware pipeline for comparing textual and market representational spaces in cryptocurrency research, together with a positive-control and power apparatus that bounds what such a comparison can detect. (2) We report a power-limited non-detection of claims–market alignment on a clean corpus. (3) We show how undetected corpus contamination can manufacture a spurious cross-sectional result, motivating content-level corpus verification as a precondition for text–market analysis.

The remainder is organised as follows. Section 2 reviews related work; Section 3 describes the data; Section 4 sets out the pipeline; Section 5 reports the results; and Sections 6–7 interpret and conclude. The extended methodology, the full robustness battery, and an extended discussion are relocated to the Supplementary Appendix.

2 Related Work

Our question sits between two readings of how narrative relates to price: the efficient-market view, on which informative narratives are quickly impounded (Fama, 1970), and Shiller's narrative economics, on which stories can decouple price from fundamentals (Shiller, 2017). Prior whitepaper research is largely cross-sectional and at issuance—technical depth and informativeness predict ICO success and post-listing returns (Thewissen et al., 2022; Momtaz, 2021; Howell et al., 2020)—whereas we test correspondence with *ongoing* market structure; notably, Suriano et al. (2025) find that whitepaper-based

clustering does not separate time-series dynamics. Methodologically we draw on Procrustes rotation (Schönemann, 1966) and Tucker's congruence coefficient (Tucker, 1951; Lorenzo-Seva and ten Berge, 2006), for which Lorenzo-Seva and ten Berge (2006) treat .85–.94 as fair agreement and $\geq$ .95 as effectively equal; we adopt a more conservative $\phi \geq 0.65$ as our own working benchmark for moderate similarity (with < 0.65 read as weak or none). The full survey—cryptocurrency narratives and sentiment, NLP in finance, factor and tensor models, and factor-comparison methods—is given in Supplementary Appendix A.

3 Data

Market data. We collect hourly OHLCV data via the Binance API (through CCXT) for 49 cryptocurrency assets spanning 1 January 2023 to 31 December 2024; the full panel is 17,543 hourly timestamps, but seven assets have shorter coverage (POL, RENDER, XMR, OCEAN, SUI, ARB, RPL; see the Table 1 note) and their cross-sectional statistics are computed over the available window. Selection follows liquidity and data-availability criteria and spans major coins (BTC, ETH) alongside a diverse set of DeFi, infrastructure, and utility tokens. Table 1 summarises the dataset.

Table 1: Dataset Summary

Dimension	**Value**
Assets (market data)	49
Assets (verified whitepapers w/ matched market data)	43
Time period	Jan 2023 – Dec 2024
Panel length (hourly)	17,543†
Market features (OHLCV)	5
Derived statistics	7
Narrative categories	10

† Panel length; 7 of 49 assets have shorter coverage (hourly rows: POL 2,630, RENDER 3,808, XMR 9,962, OCEAN 13,130, SUI 14,604, ARB 15,584, RPL 17,127). Their cross-sectional statistics use the available window; the jackknife (Supplementary Appendix D) confirms robustness to dropping individual assets, including the two most truncated (POL, RENDER).

Whitepaper corpus. We collect and *content-verify* whitepapers for 43 assets that have both an official foundational document and matched market data (AAVE, ADA, ALGO, API3, ARB, ATOM, AVAX, BTC, COMP, CRV, DOT, ETH, FIL, GRT, ICP, LINK, MKR, NEAR, SC, SOL, STORJ, UNI, XMR, ZEC, and others; the full list is in Appendix H, with corpus statistics in Supplementary Table 10). Each document is verified for word count and correct-project provenance (right-document, name-match) before inclusion, after an earlier version of the corpus was found to contain failed-download stub and wrong-document files. PDF text is extracted with sentence-level tokenisation; assets without extractable PDFs use the official markdown documentation. Intersecting the verified corpus with the market panel ($n = 49$) yields 43 common assets for alignment (Table 2); six market-panel assets are retained for the tensor calibration base but lack a usable whitepaper, and several verified whitepapers (e.g. AR, DCR, ZIL) lack matched market data and are excluded.

Table 2: Data Flow and Asset Coverage

Data Source	Assets	Notes
Whitepaper corpus (verified)	43	Content-verified documents
Market data (Binance)	49	2-year hourly OHLCV
Tensor factors (CP)	49	Rank-2 decomposition (calibration base)
NLP ∩ Market	43	Primary analysis sample
Market-only	6	Market data, no usable whitepaper

4 Methodology

The reported pipeline has three stages: (i) NLP claims extraction, (ii) market-statistics computation, and (iii) Procrustes alignment with congruence testing. We additionally build a market tensor and its CP decomposition, but—for the reasons given in the scope note below—this is *not* aligned against the claims; it serves only as a real-data base for the positive-control power calibration (Section 4.4). The tensor construction, CP/ALS decomposition, rank selection, and Tucker decomposition are set out in full in Supplementary Appendix B.

Scope of the factor representation. We do not report a factor-based alignment leg. Under the global feature-slice normalisation used to build the tensor, the CP factor matrix is degenerate for raw-scale OHLCV features: Bitcoin's level dominates the leading factor (a multi-sigma loading against a near-zero mean), so the asset-factor space mostly re-expresses Bitcoin's magnitude rather than a portable latent structure. We therefore restrict the letter's reported correspondence to the claims–statistics comparison, and retain the factor matrix only as a calibration base for the positive control (Sections 4.4 and 5.3), where the quantity of interest is the estimator's ability to recover an *injected* congruence, not the base matrix's own loadings.

4.1 NLP Claims Extraction

We classify whitepaper text with BART-large-MNLI (Lewis et al., 2020) for zero-shot classification via the HuggingFace Transformers library.[1] Documents are segmented into 500-word chunks and each chunk is scored against ten domain-relevant categories (Store of Value, Medium of Exchange, Smart Contracts, DeFi, Governance, Scalability, Privacy, Interoperability, Data Storage, Oracle Services; Supplementary Table 5), following the entailment approach of Yin et al. (2019). For text segment t and labels $\{l_1,\ldots,l_K\}$ the model returns independent per-category entailment probabilities $\sigma(s_{tk}) = (1+e^{-s_{tk}})^{-1}$, scored separately under the `multi_label` setting rather than as a softmax over categories. We row-normalise these scores into a composition and aggregate across an asset's chunks:

$$c_{nk} = \frac{1}{|T_n|} \sum_{t \in T_n} \frac{\sigma(s_{tk})}{\sum_j \sigma(s_{tj})}, \tag{1}$$

yielding the claims matrix $\mathbf{C} \in \mathbb{R}^{N \times K}$. Classifier reliability is the binding constraint on this study (Section 4.4): inter-model top-1 agreement against DeBERTa-v3 is fair ($\kappa = 0.25$; 68% at relaxed top-3), and a three-method comparison (BART-NLI, sentence embeddings, a local LLM) gives mean pairwise $r \approx 0.31$. The full validation—taxonomy, inter-model agreement, multi-method correlations, and per-category breakdown—is in Supplementary Appendix C.

[1] Model: `facebook/bart-large-mnli`, fine-tuned on MNLI (Williams et al., 2018).

Contamination-aware verification protocol. The more consequential data-quality issue is not text sparsity but contamination. In assembling the content-verified corpus we screened every document on two content-level criteria computed from the extracted text (not file metadata): a word-count threshold, which removes failed-download stub pages masquerading as whitepapers; and a project-name provenance check, which removes wrong-document files (most consequentially a GitHub-fallback retrieval that substituted the Binance Smart Chain whitepaper for Cosmos, together with wrong-document files for ADA, NEAR, and GRT). Assets failing verification were re-collected from genuine sources or dropped. This protocol is the precondition for the analysis below; Section 5 shows that omitting it manufactures a spurious cross-sectional result (further detail in Appendix I).

4.2 Market Statistics

For each asset we compute seven summary statistics and z-normalise them cross-sectionally (across assets): mean return $\bar{r}_a$; volatility σ_a; annualised Sharpe ratio $\mathrm{SR}_a = (\bar{r}_a/\sigma_a)\sqrt{252 \cdot 24}$; maximum drawdown MDD_a; average volume $\bar{V}_a$; vol-of-vol $\sigma_{\sigma,a}$ (rolling-volatility standard deviation); and price trend β_a from $P_t = \alpha + \beta t + \varepsilon$. This yields the statistics matrix $\mathbf{S} \in \mathbb{R}^{M\times 7}$. (The Sharpe annualisation factor $\sqrt{252 \cdot 24}$ follows a 252-trading-day convention; because each statistic is z-scored cross-sectionally before alignment, any positive multiplicative constant—and hence the choice between a 252-day and a 24/7 365-day crypto convention—cancels and changes no reported quantity.)

Choice of statistics. These seven quantities are an exploratory summary of cross-sectional market structure, not a theory-derived mapping from semantic categories to particular moments: we do not posit that a given narrative category should load on a specific statistic. They span the dimensions along which cryptocurrency returns are known to be organised in the asset-pricing and crypto-factor literature—realised performance (mean return, Sharpe, price trend), risk (volatility, maximum drawdown, vol-of-vol), and liquidity (average volume) (Liu et al., 2022; Dobrynskaya, 2020; Fama and French, 1993). Because no established theory links whitepaper functional claims to individual return moments, we treat the panel as a descriptive market-behaviour representation, and read a non-detection as evidence against alignment with *this summary*, not against alignment with market behaviour in general.

Collinearity and effective dimensionality. The seven statistics are not independent. Two clusters dominate: a performance cluster (mean return, Sharpe, and trend co-move, with mean return–Sharpe $r = 0.94$) and a risk cluster (volatility, maximum drawdown, and vol-of-vol, pairwise $r \approx 0.35$–0.51), with average volume the most nearly independent axis. The z-scored market matrix therefore has a participation-ratio effective rank of ≈ 3.6 (of 7; condition number ≈ 12.6, leading principal component 42% of variance), so the market space presented to the alignment is closer to a low-dimensional return/risk/liquidity factor structure than to seven independent properties. This collinearity does not spuriously inflate the reported congruence relative to its own null: the permutation test recomputes ϕ against this same matrix, so its reference distribution already absorbs the low-rank geometry, and the matched-dimension estimator (Section 4.3.1) rotates within the informative subspace. What it does do is narrow scope—the test can bear on alignment with the dominant return and risk factors but is effectively blind in the two near-null market directions (eigenvalues 0.10 and 0.02)—which we fold into the power-limited reading of the null rather than treat as a separate confound.

4.3 Procrustes Alignment and Tucker's ϕ

Definition 4.1 (Orthogonal Procrustes Problem)**.** Given matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times p}$, find orthogonal $\mathbf{Q} \in \mathbb{R}^{p \times p}$ minimising $\|\mathbf{AQ} - \mathbf{B}\|_F^2$.

Theorem 4.1 (Schönemann 1966)**.** *The optimal rotation is* $\mathbf{Q}^* = \mathbf{UV}^\top$ *where* $\mathbf{U\Sigma V}^\top = SVD(\mathbf{A}^\top \mathbf{B})$.

Proof. See Appendix F. ■

After alignment we compute the congruence coefficient between aligned columns,

$$\phi(\mathbf{x}, \mathbf{y}) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \cdot \sum_i y_i^2}}, \tag{2}$$

which for column-centred inputs equals the mean Pearson correlation across columns; the pipeline centres both matrices before rotation (standard for Procrustes), so the reported coefficient is a correlation-type congruence on centred data rather than an uncentred Tucker congruence. We report the mean absolute ϕ across columns. We adopt $|\phi| \geq 0.65$ as our conservative working benchmark for moderate similarity, reading $|\phi| < 0.65$ as weak or none.[2]

4.3.1 Dimension Handling

When the two spaces differ in column dimension we zero-pad the smaller matrix before alignment (e.g. padding the 7-dimensional statistics matrix with three zero columns to match the 10-dimensional claims matrix). This preserves all information but introduces a downward bias in ϕ: averaging seven informative market dimensions over ten slots caps the zero-padded coefficient at $7/10 = 0.70$ even against a perfectly aligned market, and deflates it by exactly $10/7$ relative to the same rotation's mean over its seven informative columns (0.223 versus 0.319); the matched-dimension estimate (0.303) is a separate SVD-reduced rotation that lands close to that un-padded value, approximately $1.4\times$ (precisely $1.36\times$) the zero-padded coefficient. Because that ceiling coincides with the strong-alignment threshold, the zero-padded estimator can serve only as a conservative display floor; it cannot, by construction, bear on whether $\phi \geq 0.70$. We therefore take as our *primary* estimate a matched-dimension variant that SVD-reduces the higher-dimensional matrix with no padding (Supplementary Appendix D), whose congruence against a perfectly aligned market reaches 1.0—so the strong-alignment threshold is geometrically reachable and a rejection there has content—and report the zero-padded coefficient ($\phi = 0.223$) only as a lower bound.

4.4 Power and Inference

With $n = 43$ common entities, statistical power to detect alignment is limited. A positive-control Monte Carlo (500 iterations per effect size, 200 permutations each; seed 20260627; Section 5.3) injects a known cross-domain congruence of magnitude ϕ into a calibration base and pushes it through the unmodified pipeline. The floor summarised here is the *matched-geometry* variant that uses the 43×7 statistics matrix as the base, mirroring the primary claims–statistics test's dimension; the factor-matrix carrier and the full derivation (including the higher factor-base floors) are in Section 5.3. We report power under two instruments: an *ideal* instrument (clean signal, isolating estimator sensitivity) and a *realistic* instrument whose claims side is attenuated by the estimator-relevant classifier reliability—the within-category, between-asset agreement (mean $\rho \approx 0.57$, evaluated at the nearest grid reliability $\rho = 0.50$), not the pooled cross-cell correlation:

[2] These thresholds were developed for factor solutions from similar data; applying them across heterogeneous NLP-derived and market-derived spaces extends beyond their original validation context, so we use them as rough benchmarks.

True ϕ	Power (ideal)	Power (realistic, $\rho \approx 0.50$)
0.20	44%	13%
0.30	83%	42%
0.50	100%	96%
0.65	100%	100%
0.80	100%	100%

The estimator itself is sensitive but capable: under an ideal instrument the matched-geometry minimum detectable effect (80% power, $\alpha = 0.05$) is $\phi \approx 0.29$ (the factor-matrix carrier's ideal floor is the higher $\phi \approx 0.40$; Section 5.3). The binding constraint is the noisy claims instrument—but the relevant reliability is not the pooled cross-cell correlation. The mean of the three pairwise inter-method correlations, pooled over all asset×category cells, is $\rho \approx 0.31$ (BART–embedding 0.070, BART–LLM 0.475, embedding–LLM 0.395; Supplementary Table 6), but this pooled figure conflates category-level mean and scale disagreement that the column-centred estimator removes before computing congruence. The estimator-relevant quantity is the *within-category, between-asset* agreement, which the same outputs put at mean $\rho \approx 0.57$ (category range 0.21–0.82). On the matched-geometry positive control at this reliability the minimum detectable effect is $\phi \approx 0.44$ (evaluated at $\rho = 0.50$; it rises to $\phi \approx 0.53$ at the pooled $\rho = 0.31$, and at the near-zero BART–embedding pair $\rho = 0.07$ detection is effectively blind within the simulated grid, power $\approx 46\%$ even at $\phi = 0.8$). We take $\phi \approx 0.44$ as the working floor and read it as an upper bound on what the design can resolve: the study excludes not only strong but *moderate* alignment ($\phi \geq 0.44$ at roughly 80% power), while remaining unable to separate weak alignment ($\phi \approx 0.3$) from none. The near-zero BART–embedding agreement (0.07) remains the binding concern: two of the three instruments barely covary, which suggests they may not be measuring the same construct. The matched-dimension estimate ($\phi = 0.303$) is therefore inconsistent with moderate-or-stronger alignment, not evidence that whitepaper claims and market structure are unrelated.

Significance is assessed by a one-sided permutation test (permute rows of $\mathbf{B}$, recompute ϕ; $B = 1000$; $p = \frac{1}{B}\sum_b \mathbf{1}[\phi^{(b)} \geq \phi^*]$, testing H_0: $\phi \leq \phi_{\text{random}}$). We also computed percentile bootstrap CIs but do not report them as headline quantities: small-sample Procrustes resampling induces a documented upward bias whose lower bound can exceed the point estimate (Supplementary Appendix D.12). All stochastic procedures use fixed seeds (42 for CP-ALS, tensor operations, and permutation tests); implementation is Python 3.11+ with NumPy, SciPy, TensorLy, scikit-learn, and HuggingFace Transformers. Code and data: `https://github.com/studiofarzulla/whitepaper-claims`.

5 Results

We report the three load-bearing results—the bounded null on the claims–statistics leg, the collapse of the apparent entity-impact ordering under verification, and the positive control—and point to the Supplementary Appendix for the claims matrix, temporal stability, feature importance, and the full robustness battery (subsample stability, Bitcoin sensitivity, alternative metrics, jackknife, market-cap control, multiple-testing), all of which corroborate the null.

5.1 Primary Alignment: A Bounded Null

Table 3 reports the claims–statistics alignment under the conservative zero-padded estimator.

Table 3: Alignment Test Result — Zero-Padded Tucker ϕ (Conservative Variant, $n = 43$)

Comparison	$\boldsymbol{\phi}$	**p-value**	**Interpretation**
Claims–Statistics	0.223	0.433	Weak

Note: the p-value is a one-sided permutation test ($B = 1000$). We omit bootstrap confidence intervals from the headline table: small-sample Procrustes resampling produces a documented upward bias (Supplementary Appendix D.12) that renders the percentile interval incoherent—its lower bound can exceed the point estimate. The zero-padded coefficient reported here is a mean absolute *congruence with a one-sided test; the matched-dimension coefficient ($\phi = 0.303$) is a mean* signed *congruence with a two-sided test—the two headline estimates use different statistics, and both are non-significant.*

Because the claims (10D) and statistics (7D) spaces differ in dimension, the zero-padded coefficient is mechanically deflated; Table 3 reports it as a conservative lower bound. Our *primary*, matched-dimension estimate (SVD-reduction, no padding; Supplementary Appendix D) is claims–statistics $\phi = 0.303$ ($p = 0.365$). Under either estimator the comparison is weak ($\phi \leq 0.303$) and non-significant, well below the 0.65 moderate-similarity threshold. At this sample size we read this as a non-detection—whatever cross-sectional market structure exists is not registered in whitepaper content—while noting that a weak sub-threshold effect cannot be excluded (Section 4.4). The positive control (Section 5.3) confirms the estimator recovers an injected cross-domain congruence, so this non-detection reflects limited power, not an insensitive pipeline.

5.2 Entity-Level Analysis: An Ordering That Did Not Survive

The sharpest illustration of why corpus verification matters is the entity-level analysis. Figure 1 shows leave-one-out entity impact before and after content verification.

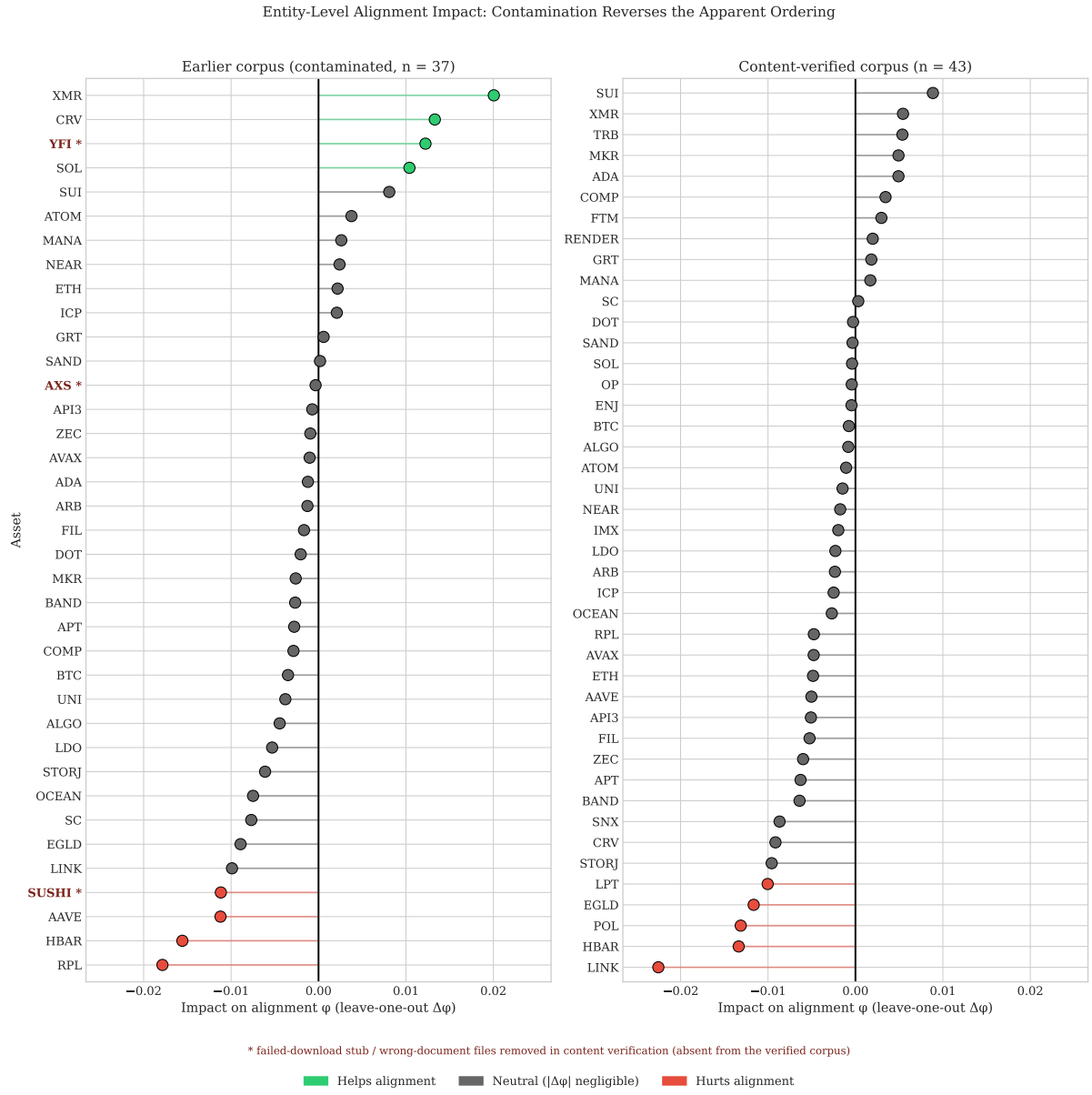

Figure 1: Entity-level alignment impact, before and after content verification, on a common horizontal scale. *Left:* the earlier, contaminated corpus ($n = 37$), on which several tokens appear to *help* alignment—including YFI (starred), subsequently found to be a contaminated file and absent from the verified corpus—producing an apparent specialised-versus-infrastructure ordering. *Right:* the content-verified corpus ($n = 43$), on which no token registers as helping: leave-one-out $\Delta\phi$ is negligible across the corpus and the apparent ordering does not survive. Starred labels mark failed-download stub or wrong-document files removed in verification. Points are coloured by interpretation (helps / neutral / hurts). The verified corpus is larger than the contaminated one ($n = 43$ versus 37), not smaller, because the two panels are different corpus snapshots: between them the corpus was expanded from an initial 24-asset set and the mis-retrieved documents were re-collected from genuine sources, adding more verified whitepapers than screening removed—so the rise in n reflects net corpus growth, not a relaxation of the verification criteria.

Table 4 gives the top and bottom contributors on the verified corpus.

Table 4: Entity Impact Analysis (Top/Bottom)[3]

Asset	Impact	Interpretation
SUI	$+0.009$	Marginal
XMR	$+0.005$	Marginal
TRB	$+0.005$	Marginal
MKR	$+0.005$	Marginal
EGLD	-0.012	Hurts alignment
POL	-0.013	Hurts alignment
HBAR	-0.013	Hurts alignment
LINK	-0.022	Hurts alignment

On the content-verified corpus no coherent specialised-versus-infrastructure pattern emerges: 38 of 43 entities have negligible leave-one-out impact ($|\Delta\phi| < 0.01$), the largest positive contributors are weak and heterogeneous (SUI $+0.009$, XMR $+0.005$, TRB $+0.005$), and the largest negative contributors are broad L1/infrastructure tokens (LINK -0.022, HBAR -0.013, POL -0.013). An earlier version of this corpus reported a striking ordering—privacy and DeFi tokens appearing to align while broad infrastructure tokens detracted—but several of the tokens defining that ordering were failed-download stub or wrong-document files rather than genuine whitepapers; once the contaminated documents are removed or re-collected from verified sources, the ordering does not survive. Bitcoin, despite its exceptional cross-sectional leverage, has negligible impact (-0.001): removing it leaves the weak, non-significant correspondence essentially unchanged. We therefore treat entity-level heterogeneity as uninformative here, consistent with the overall null and as a cautionary illustration of how corpus contamination can manufacture an apparent cross-sectional signal.

5.3 Positive Control: Cross-Domain Detection Sensitivity

To confirm that the pipeline can detect *cross-domain* congruence, we run a positive-control simulation. The carrier of the injected signal is the real clean-43 CP asset-factor matrix—a 43×2 matrix (rank $R = 2$; Section 4), used here only as a calibration base. For a grid of true congruence values $\phi \in \{0.20, 0.30, 0.50, 0.65, 0.80\}$ we corrupt each of its two columns with calibrated Gaussian noise so that the per-column congruence with the clean column equals the target ϕ, then embed the resulting two-column signal into a two-dimensional subspace of the ten-dimensional claims space via a random orthonormal map ($\mathbf{U} \in \mathbb{R}^{10\times 2}$, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_2$). This is an isometric embedding of the rank-two factor signal into the claims space—not a square rotation of one 43×10 matrix into another—and yields a synthetic 43×10 claims-like matrix carrying exactly the injected two-dimensional structure plus its calibrated noise. We pass it through the unmodified Procrustes and permutation pipeline (500 replications $\times$ 200 permutations per cell; seed 20260627) and measure recovery on the two genuine factor axes. Recovered congruence on those two axes tracks the injected target almost exactly (targets $0.20/0.50/0.80$ recover to $0.234/0.501/0.806$), validating the data-generating process. Because the control injects and then recovers a *known* congruence on top of the calibration base, the base's own factor structure—whose leading component is Bitcoin-dominated—does not affect the recovered power; only the injected signal does. This establishes estimator sensitivity *conditional on* a signal being present in the extracted representation; it is silent on whether the text instrument captures the price-relevant construct in the first place—a

[3] Selection criterion: top 4 and bottom 4 entities by absolute impact magnitude from leave-one-out analysis.

question of validity, not power (Section 6). As a check that this calibration transfers to the headline estimator, re-running the control with the real 43×7 statistics matrix as the base and the matched-dimension estimator—synthetic claims reduced $10 \to 7$ before congruence, mirroring the primary test's geometry exactly—yields a comparable, slightly lower detection floor (ideal minimum detectable effect $\phi \approx 0.29$; mean-reliability floor $\phi \approx 0.53$; `positive_control_matched_statsbase.json`), so the figures reported below are, if anything, conservative. The resulting power curve for this primary control shows the estimator is sensitive to genuine cross-domain structure—under an ideal instrument the minimum detectable effect (80% power) is $\phi \approx 0.40$—and that the binding constraint is the low reliability of the claims instrument: at the estimator-relevant within-category reliability ($\rho \approx 0.50$) the practical minimum detectable effect is $\phi \approx 0.44$, rising to $\phi \approx 0.53$ at the pooled mean of the three heterogeneous pairwise correlations ($\rho \approx 0.31$; Section 4.4). The observed claims–statistics congruence lies far below this floor, so the null is consistent with a true null within the limits of detectable effect size, though a weak sub-threshold effect cannot be excluded.

5.4 Robustness

The non-detection is stable across the full battery, reported in Supplementary Appendix D: the claims matrix is near-homogeneous and recovers no intuitive token archetype (Supplementary Figure 3); alignment is weak in every rolling window ($\phi = 0.163 \pm 0.028$; Supplementary Table 7); no single semantic category moves ϕ by more than 0.02 (Supplementary Table 8, Figure 4); subsample resampling ($\phi = 0.254 \pm 0.022$) and jackknife leave the estimate firmly weak; four methodologically distinct alignment measures (RV, distance correlation, CCA, PLS) are all non-significant (Supplementary Table 9); residualising on market capitalisation leaves the result unchanged (partial $\phi \approx 0.26$); and because no comparison clears $\alpha = 0.05$, any multiple-testing correction leaves the conclusion intact.

6 Discussion

Interpreting the null. Our central result is a non-detection: at $n = 43$ we find no significant claims–market alignment, but the design is underpowered (realistic minimum detectable effect $\phi \approx 0.44$ at the estimator-relevant reliability) to separate weak alignment from none. Our results are inconsistent with strong alignment ($\phi \geq 0.70$) but cannot rule out a weak ($\phi \approx 0.3$) effect. Three readings are consistent with the data: whitepapers may state aspirational narratives that projects later pivot away from (Bitcoin's "peer-to-peer electronic cash" framing diverged sharply from its "digital gold" market reality); market behaviour may be driven by factors orthogonal to functional claims—speculation, liquidity, Bitcoin co-movement, macro (Liu and Tsyvinski, 2021)—that swamp project-specific narrative; or the zero-shot pipeline may simply fail to capture price-relevant narrative nuance.

Contamination as a cautionary tale. We initially observed what looked like a structural distinction—specialised tokens with clear niches aligning while broad infrastructure tokens detracted—but this ordering was an artefact of corpus contamination: several of the tokens defining it were failed-download stub or wrong-document files, and the split dissolves once they are removed or re-collected. On the verified corpus the entity-level impacts are uniformly small ($|\Delta\phi| \leq 0.022$) with no coherent niche-versus-infrastructure structure. We therefore advance no cross-sectional pricing hypothesis. Content verification is itself elementary data hygiene rather than a methodological novelty; what the episode contributes is a concrete demonstration of how much a plausible cross-sectional result can rest on it—here roughly a quarter of one corpus build—and thus why content-level corpus verification (word counts, project-name checks, right-document confirmation) should be treated as a precondition for, not an optional refinement of, cross-sectional text–market analysis. (Extended theoretical and practical implications are in

Supplementary Appendix E.)

Limitations. Three limitations bind the interpretation; further caveats are in Supplementary Appendix E.

- **Limited power is the binding constraint.** With $n = 43$ entities and a low-reliability text instrument, the realistic minimum detectable effect is $\phi \approx 0.44$ at the estimator-relevant reliability (Section 4.4); our results are inconsistent with strong alignment ($\phi \geq 0.70$), for which the design is adequately powered (>80%), but cannot distinguish weak alignment from none. The result is absence of evidence for alignment, not evidence of its absence. Expanding to 50+ projects would be needed for adequate power at moderate effects and for sector-level subsample analysis.

- **Construct validity of the narrative instrument.** Inter-method agreement is modest and uneven ($\kappa = 0.25$; pairwise ρ from 0.07 for BART–embedding to 0.48 for BART–LLM, mean ≈ 0.31). This bears not only on reliability—noise that merely attenuates a true signal—but on validity: off-the-shelf zero-shot classification may not measure the *price-relevant* narrative content the test requires. The near-zero BART–embedding agreement is the sharpest version of the problem, since two of the three instruments barely covary and so may not be tracking the same construct at all. Two further facts sharpen the concern. First, every reliability figure we report is inter-*method* (model-versus-model) agreement; we have no human-adjudicated ground truth for the category labels, so the instrument's absolute validity—as opposed to the mere concordance of two models—is unestablished. Second, the classifier fails an elementary face-validity check: it does not recover Bitcoin's central store-of-value narrative (which ranks only sixth, at 7.0%, behind a modal Medium-of-Exchange weight; Appendix D), and more broadly compresses heterogeneous technical prose toward a generic monetary register. A weak measured alignment is thus consistent with either a genuinely weak relationship or an instrument that misses the construct, and our design cannot separate the two.

- **Temporal-coverage mismatch.** The corpus is mostly founding-era text (Supplementary Table 10; several documents predate 2018), whereas market structure is measured over 2023–2024. Treating inception narratives and contemporaneous market behaviour as comparable can attenuate alignment and cannot be cleanly separated from the validity issue; resolving it requires a time-stamped, multi-period narrative corpus matched to market windows.

Practical implications. We are cautious about prescriptions from an underpowered non-detection. A study that can exclude only strong alignment cannot license claims that whitepaper analysis "offers limited value" to investors, that messaging is secondary for project teams, or that a narrative–market disconnect complicates disclosure-based regulation. At most our results offer a weak prior—narrative-classification strategies (e.g. "DeFi basket," "Layer 1 portfolio") should not be assumed to capture return differentials without independent evidence—and motivate the larger, feature-enriched study any firmer recommendation would require.

7 Conclusions

We investigated whether cryptocurrency whitepaper claims are reflected in market structure, combining NLP claims extraction, cross-sectional market-statistics construction, and Procrustes alignment under a contamination-aware verification protocol. The result is layered. The claims–market comparison is weak and non-significant (dimension-matched $\phi = 0.303$), and a positive-control and power analysis shows the design is inconsistent with strong alignment ($\phi \geq 0.70$) but cannot adjudicate weak alignment

from none—so we report a power-limited non-detection, not a demonstrated decoupling of narrative from markets. Separately, an apparent entity-level ordering reported for an earlier corpus did not survive content verification; we document it as a corpus-contamination artefact and a methodological caution.

The paper's contribution is therefore a method plus a cautionary tale: a reproducible, content-verified pipeline for comparing textual and market representational spaces, with a positive-control power apparatus that bounds what the comparison can detect; a power-limited empirical null on a clean corpus; and a demonstration that undetected contamination can manufacture a spurious cross-sectional result. Whether the weak measured alignment reflects a genuine narrative–market disconnect, an under-powered design, or a low-validity text instrument remains open: our data are inconsistent with strong alignment but cannot adjudicate among these accounts. The natural next steps—a 50+-project corpus, domain-adapted classifiers, crypto-native market features, and a time-stamped narrative corpus matched to market windows (Supplementary Appendix E)—would raise power and validity enough to make the bounded null decisive in either direction.

Acknowledgements

Portions of this manuscript were drafted and revised with assistance from Claude (Anthropic). The author retains full responsibility for all intellectual content, analytical decisions, and interpretive claims. This work was conducted as part of the Adversarial Systems Research programme at Dissensus. Working papers and replication materials are available through the Adversarial Systems & Complexity Research Initiative (ASCRI). Comments and correspondence are welcome at murad@dissensus.ai.

Declarations

Conflict of Interest. The author declares no competing interests.

Funding. This research received no external funding. Computational resources were provided by King's College London.

Data Availability. Cryptocurrency market data obtained via the Binance exchange API through CCXT (publicly available). Whitepaper corpus collected from official project documentation (publicly available). Processed datasets and NLP classification outputs available at `https://github.com/studiofarzulla/whitepaper-claims`.

Code Availability. Full replication code available at `https://github.com/studiofarzulla/whitepaper-claims` under CC BY 4.0.

AI Assistance. Claude (Anthropic) assisted with manuscript drafting, code review, and statistical exposition. All research design, data analysis, and interpretation are solely the author's.

Author Contributions. Sole author.

References

Saman Adhami, Giancarlo Giudici, and Stefano Martinazzi. Why do businesses go crypto? an empirical analysis of initial coin offerings. *Journal of Economics and Business*, 100:64–75, 2018. doi: 10.1016/j.jeconbus.2018.04.001.

Lennart Ante. How Elon Musk's Twitter activity moves cryptocurrency markets. *Technological Forecasting and Social Change*, 186:122112, 2023. doi: 10.1016/j.techfore.2022.122112.

Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

Tomaso Aste. Cryptocurrency market structure: connecting emotions and economics. *Digital Finance*, 1:5–21, 2019. doi: 10.1007/s42521-019-00008-9.

Tomaso Aste. Information filtering networks: Theoretical foundations, generative methodologies, and real-world applications. *arXiv preprint arXiv:2505.03812*, 2025.

Malcolm Baker and Jeffrey Wurgler. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680, 2006. doi: 10.1111/j.1540-6261.2006.00885.x.

Malcolm Baker and Jeffrey Wurgler. Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2):129–152, 2007. doi: 10.1257/jep.21.2.129.

Nicholas Barberis and Richard Thaler. A survey of behavioral finance. *Handbook of the Economics of Finance*, 1:1053–1128, 2003. doi: 10.1016/S1574-0102(03)01027-6.

Silvia Bartolucci, Giuseppe Destefanis, Marco Ortu, Nicola Uras, Michele Marchesi, and Roberto Tonelli. The butterfly "affect": impact of development practices on cryptocurrency prices. *EPJ Data Science*, 9(1):21, 2020. doi: 10.1140/epjds/s13688-020-00239-6.

Siddharth Bhambhwani, Stefanos Delikouras, and George M Korniotis. Do fundamentals drive cryptocurrency prices? *SSRN Electronic Journal*, 2019. doi: 10.2139/ssrn.3342842.

Daniele Bianchi and Mykola Babiak. A factor model for cryptocurrency returns. *SSRN Electronic Journal*, 2021. doi: 10.2139/ssrn.3935934.

Antonio Briola and Tomaso Aste. Dependency structures in cryptocurrency market from high to low frequency. *Entropy*, 24(11):1548, 2022. doi: 10.3390/e24111548.

Antonio Briola, David Vidal-Tomás, Yuanrong Wang, and Tomaso Aste. Anatomy of a stablecoin's failure: the Terra-Luna case. *Finance Research Letters*, 51:103358, 2022. doi: 10.1016/j.frl.2022.103358.

Antonio Briola, Marwin Schmidt, Fabio Caccioli, Carlos Ros Perez, James Singleton, Christian Michler, and Tomaso Aste. Graph regularized PCA. *arXiv preprint arXiv:2601.10199*, 2026.

Frank B Brokken. Orthogonal Procrustes rotation maximizing congruence. *Psychometrika*, 48(3):343–352, 1983. doi: 10.1007/BF02293679.

Andrew Burnie, Emine Yilmaz, and Tomaso Aste. Analysing social media forums to discover potential causes of phasic shifts in cryptocurrency price series. *Frontiers in Blockchain*, 3:610231, 2020. doi: 10.3389/fbloc.2020.00001.

Fabio Caccioli, Paolo Barucca, and Teruyoshi Kobayashi. Network models of financial systemic risk: A review. *Journal of Computational Social Science*, 1(1):81–114, 2018. doi: 10.1007/s42001-017-0008-3.

Rong Chen, Dan Yang, and Cun-Hui Zhang. Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117(537):94–116, 2022. doi: 10.1080/01621459.2021.1912757.

Zheshi Chen, Chunhong Li, and Wenjun Sun. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 365:112395, 2019. doi: 10.1016/j.cam.2019.112395.

Victoria Dobrynskaya. Is downside risk priced in cryptocurrency market? *SSRN Electronic Journal*, 2020. doi: 10.2139/ssrn.3693837.

Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970. doi: 10.2307/2325486.

Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993. doi: 10.1016/0304-405X(93)90023-5.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(4):603–680, 2013. doi: 10.1111/rssb.12016.

Murad Farzulla. Do cryptocurrency markets differentiate infrastructure from regulatory shocks? a multi-moment event study with dependence-robust inference. arXiv preprint arXiv:2602.07046, 2026. doi: 10.5281/zenodo.18099608.

Christian Fisch. Initial coin offerings (ICOs) to finance new ventures. *Journal of Business Venturing*, 34 (1):1–22, 2019. doi: 10.1016/j.jbusvent.2018.09.007.

David Florysiak and Alexander Schandlbauer. Experts or charlatans? ICO analysts and white paper informativeness. *Journal of Banking & Finance*, 139:106476, 2022. doi: 10.1016/j.jbankfin.2022.106476.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740.

Yuefeng Han, Dan Yang, Cun-Hui Zhang, and Rong Chen. CP factor model for dynamic tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):713–740, 2024. doi: 10.1093/jrsssb/qkae036.

Richard A Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

Ozkan Haykir and İbrahim Yağlı. Speculative bubbles and herding in cryptocurrencies. *Financial Innovation*, 8(1):1–33, 2022. doi: 10.1186/s40854-022-00383-0.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021.

Sabrina T Howell, Marina Niessner, and David Yermack. Initial coin offerings: Financing growth with cryptocurrency token sales. *The Review of Financial Studies*, 33(9):3925–3974, 2020. doi: 10.1093/rfs/hhz131.

Allen H. Huang, Hui Wang, and Yi Yang. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023. doi: 10.1111/1911-3846.12832.

Colm Kearney and Sha Liu. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185, 2014. doi: 10.1016/j.irfa.2014.02.006.

Zeynep Keskin and Tomaso Aste. Information-theoretic measures for nonlinear causality detection: application to social media sentiment and cryptocurrency prices. *Royal Society Open Science*, 7(9): 200863, 2020. doi: 10.1098/rsos.200863.

Kemal Kirtac and Guido Germano. Large language models in finance: what is financial sentiment? *arXiv preprint arXiv:2503.03612*, 2025.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Review*, 51(3): 455–500, 2009. doi: 10.1137/07070111X.

Bruce Korth and Ledyard R Tucker. The distribution of chance congruence coefficients from simulated data. *Psychometrika*, 40(3):361–372, 1975. doi: 10.1007/BF02291763.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020. doi: 10.18653/v1/2020.acl-main.703.

Yukun Liu and Aleh Tsyvinski. Risks and returns of cryptocurrency. *The Review of Financial Studies*, 34(6):2689–2727, 2021. doi: 10.1093/rfs/hhaa113.

Yukun Liu, Aleh Tsyvinski, and Xi Wu. Common risk factors in cryptocurrency. *The Journal of Finance*, 77(2):1133–1177, 2022. doi: 10.1111/jofi.13119.

Giacomo Livan, Simone Alfarano, and Enrico Scalas. Fine structure of spectral properties for random correlation matrices: an application to financial markets. *Physical Review E*, 84(1):016113, 2011. doi: 10.1103/PhysRevE.84.016113.

Urbano Lorenzo-Seva and Jos MF ten Berge. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2):57–64, 2006. doi: 10.1027/1614-2241.2.2.57.

Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011. doi: 10.1111/j.1540-6261.2010.01625.x.

Tim Loughran and Bill McDonald. Textual analysis in finance. *Annual Review of Financial Economics*, 12:357–375, 2020. doi: 10.1146/annurev-financial-012820-032249.

Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T Chitkushev, and Dimitar Trajanov. Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, 8:131258–131275, 2020. doi: 10.1109/ACCESS.2020.3009626.

Paul P Momtaz. Entrepreneurial finance and moral hazard: Evidence from token offerings. *Journal of Business Venturing*, 36(5):106180, 2021. doi: 10.1016/j.jbusvent.2020.106001.

Giuseppe Pappalardo, Tiziana Di Matteo, Guido Caldarelli, and Tomaso Aste. Blockchain inefficiency in the Bitcoin peers network. *EPJ Data Science*, 7(1):1–14, 2018. doi: 10.1140/epjds/s13688-018-0159-3.

Sampo V Paunonen. On chance and factor congruence following orthogonal Procrustes rotation. *Educational and Psychological Measurement*, 57(1):33–59, 1997. doi: 10.1177/0013164497057001003.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Paul Robert and Yves Escoufier. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25(3):257–265, 1976. doi: 10.2307/2347233.

Shadi Samieifar and Dirk G Baur. Read me if you can! an analysis of ICO white papers. *Finance Research Letters*, 38:101427, 2021. doi: 10.1016/j.frl.2020.101427.

Peter H Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31 (1):1–10, 1966. doi: 10.1007/BF02289451.

Robert J Shiller. Narrative economics. *American Economic Review*, 107(4):967–1004, 2017. doi: 10.1257/aer.107.4.967.

M. Suriano, L. F. Caram, C. Caiafa, Hernán Daniel Merlino, and O. A. Rosso. Information theory quantifiers in cryptocurrency time series analysis. *Entropy*, 27(4):450, 2025. doi: 10.3390/e27040450.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007. doi: 10.1214/009053607000000505.

Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007. doi: 10.1111/j.1540-6261.2007.01232.x.

James Thewissen, Prabal Shrestha, Wouter Torsin, and Anna M Pastwa. Unpacking the black box of ICO white papers: A topic modeling approach. *Journal of Corporate Finance*, 75:102225, 2022. doi: 10.1016/j.jcorpfin.2022.102225.

Ledyard R Tucker. A method for synthesis of factor analysis studies. *Personnel Research Section Report*, (984), 1951.

David Vidal-Tomás, Antonio Briola, and Tomaso Aste. FTX's downfall and Binance's consolidation: The fragility of centralized digital finance. *Physica A: Statistical Mechanics and its Applications*, 625:129044, 2023. doi: 10.1016/j.physa.2023.129044.

Di Wang, Yao Zheng, Heng Lian, and Guodong Li. High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, 117(539):1338–1356, 2022. doi: 10.1080/01621459.2020.1855183.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1101.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1404.

Supplementary Appendix

The following material relocates, in full, the extended methodology, NLP validation, robustness battery, and extended discussion referenced in the main text. No analyses or results have been removed; the main text retains only the contamination-aware pipeline and protocol, the collapse of the apparent entity-impact ordering under verification, the positive-control/power bound, and the bounded null on the claims–statistics leg.

A Extended Related Work

A.1 Cryptocurrency Narratives and Sentiment

Research on cryptocurrency narratives spans social media, whitepaper studies, and sentiment measurement. Chen et al. (2019) predict Bitcoin price movements from engineered sample dimensions; Ante (2023) find that Elon Musk's tweets generate abnormal returns for mentioned coins; Haykir and Yağlı (2022) document narrative-driven herding; and Liu and Tsyvinski (2021) establish crypto-specific momentum, size, and market factors. Whitepaper analysis has grown as a quality signal: Howell et al. (2020), Fisch (2019), and Adhami et al. (2018) link technical depth and design to ICO success, and more sophisticated NLP has followed—Thewissen et al. (2022) topic-model 5,210 whitepapers, Momtaz (2021) shows issuers systematically exaggerate claims, Samieifar and Baur (2021) relate length and complexity to funds raised, and Florysiak and Schandlbauer (2022) find post-listing returns better predicted by whitepaper content than by analyst ratings. These studies, however, focus on prediction *at issuance* rather than ongoing alignment between narrative and market behaviour. Indeed Suriano et al. (2025) find that clustering coins by whitepaper content yields no significant difference in time-series dynamics, and high-profile failures (Briola et al., 2022; Vidal-Tomás et al., 2023) illustrate the gap between elaborate specifications and realised outcomes. Our study tests whether claims align with ongoing market *structure*. The theoretical frame is narrative economics (Shiller, 2017) and the behavioural-finance literature on sentiment and mispricing (Baker and Wurgler, 2006, 2007; Tetlock, 2007; Keskin and Aste, 2020).

A.2 NLP in Finance

Applying NLP to financial text is now standard (Loughran and McDonald, 2020): general dictionaries misclassify financial language (Loughran and McDonald, 2011; Kearney and Liu, 2014), motivating domain models such as FinBERT (Araci, 2019; Huang et al., 2023) and transformer benchmarks (Mishev et al., 2020). Zero-shot classification (Lewis et al., 2020) allows categorisation without labelled training data, and developer-communication sentiment predicts crypto prices (Bartolucci et al., 2020). Beyond sentiment, distinctive microstructure—24/7 trading, fragmentation, Bitcoin co-movement—may dominate narrative effects (Pappalardo et al., 2018), and infrastructure shocks generate larger volatility responses than regulatory ones (Farzulla, 2026).

A.3 Factor Comparison Methods

Comparing structures across spaces requires controlling for rotational indeterminacy. Procrustes rotation (Schönemann, 1966) finds the optimal orthogonal alignment, Brokken (1983) maximises congruence directly, and Tucker's ϕ (Tucker, 1951; Lorenzo-Seva and ten Berge, 2006) measures aligned-factor similarity, with chance distributions established by Korth and Tucker (1975) and Paunonen (1997). Lorenzo-Seva and ten Berge (2006) give interpretation thresholds ($|\phi| \geq 0.95$ equivalence, ≥ 0.85 fair, ≥ 0.65 moderate, < 0.65 weak/none) that we use as rough benchmarks throughout.

A.4 Factor Models in Cryptocurrency

Traditional asset pricing employs factor models to explain cross-sectional return variation. Fama (1970) established the theoretical foundation for efficient markets and factor-based returns; Fama and French (1993) introduced the three-factor model for equities, with analogous developments in cryptocurrency emerging more recently. Livan et al. (2011) demonstrate how random matrix theory can distinguish signal from noise in financial correlation matrices—a perspective we extend to narrative-factor comparisons. Caccioli et al. (2018) review network-based approaches to financial systemic risk, complementing factor-based perspectives with topological analysis, and Aste (2025) establishes the theoretical and generative foundations for information filtering networks, offering principled methods for extracting sparse dependency structures from high-dimensional financial data.

Liu and Tsyvinski (2021) document that cryptocurrency returns load on common factors explaining substantial cross-sectional variation analogous to Fama-French factors; Liu et al. (2022) subsequently formalise a three-factor model—market, size, and momentum—for the cross-section of cryptocurrency returns. Bianchi and Babiak (2021) apply Instrumented PCA to show that time-varying factor loadings outperform observable risk factors; Dobrynskaya (2020) extends crypto CAPM with a downside-beta factor across 1,700 coins; and Bhambhwani et al. (2019) introduce blockchain-native factors—computing power, network size—as procyclical pricing factors with positive risk premia. Such systematic factors—market co-movement, size, liquidity—may dominate any narrative-based signal, consistent with our non-detection of a claims–market correspondence.

Multi-way data in finance. Financial data naturally exhibits multi-way structure: assets $\times$ time $\times$ features. While matrix methods (PCA, factor analysis) collapse this structure, tensor decomposition preserves it; this literature motivates the market-structure representation we use as a calibration target for the power analysis (Section 4.4).

A.5 Tensor Methods in Finance

Tensor decomposition provides a natural framework for multi-way financial data. Kolda and Bader (2009) review tensor decomposition methods, establishing the foundations for CP and Tucker decompo-

sition. Chen et al. (2022) develop tensor factor models for high-dimensional time series (TIPUP/TOPUP estimators) with finance applications; Wang et al. (2022) apply Tucker decomposition to high-dimensional vector autoregression; Han et al. (2024) develop CP factor models for dynamic tensors with uncorrelated latent factors applicable to asset pricing; Fan et al. (2013) introduce the POET estimator for high-dimensional covariance with factor structure; and Briola et al. (2026) regularise principal components with network topology. CP (CANDECOMP/PARAFAC) decomposition decomposes a tensor into rank-one components, extracting interpretable latent factors (Harshman, 1970); for market data structured as (time $\times$ asset $\times$ feature) it yields asset-level loadings analogous to PCA but preserving multi-way structure, with Tucker decomposition an alternative offering mode-specific ranks and a core tensor.

B Tensor Construction and Decomposition (Calibration Base)

The market tensor and its CP decomposition are used *only* as a real-data base for the positive-control power calibration (Sections 4.4, 5.3); for the reasons given in the scope note in Section 4, no factor-based alignment leg is reported.

B.1 Tensor Construction

Definition B.1 (Market Tensor)**.** A market tensor $\mathscr{X} \in \mathbb{R}^{T\times V\times A\times F}$ is a 4-way array with modes:

- Time ($T = 17{,}543$ hourly timestamps)
- Venue ($V = 1$, Binance)
- Asset ($A = 49$ cryptocurrencies)
- Feature ($F = 5$, OHLCV)

With a single venue, the effective structure is 3-way: $\mathscr{X} \in \mathbb{R}^{T\times A\times F}$. Each entry x_{taf} represents the value of feature f for asset a at time t. Prior to decomposition, we z-normalise each feature slice across both assets and time (i.e., each $\mathscr{X}_{::f}$ has zero mean and unit variance), ensuring that scale differences across OHLCV features do not dominate the factor structure.

B.2 Tensor Decomposition

Definition B.2 (CP Decomposition)**.** The CANDECOMP/PARAFAC (CP) decomposition approximates a tensor as a sum of rank-one tensors:

$$\mathscr{X} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{w}_r \tag{3}$$

where $\circ$ denotes outer product, λ_r are weights, and $\mathbf{a}_r \in \mathbb{R}^T$, $\mathbf{b}_r \in \mathbb{R}^A$, $\mathbf{w}_r \in \mathbb{R}^F$ are mode-specific factor vectors.

The factor matrices are:

$$\mathbf{A} = [\mathbf{a}_1 | \cdots | \mathbf{a}_R] \in \mathbb{R}^{T\times R} \quad \text{(time factors)} \tag{4}$$

$$\mathbf{B} = [\mathbf{b}_1 | \cdots | \mathbf{b}_R] \in \mathbb{R}^{A\times R} \quad \text{(asset factors)} \tag{5}$$

$$\mathbf{W} = [\mathbf{w}_1 | \cdots | \mathbf{w}_R] \in \mathbb{R}^{F\times R} \quad \text{(feature factors)} \tag{6}$$

The asset factor matrix $\mathbf{B} \in \mathbb{R}^{A\times R}$ (the asset mode has $A = 49$) holds the latent asset loadings; in this letter it serves only as the calibration base for the positive-control power analysis (Section 4.4). We write the feature-factor matrix as $\mathbf{W}$ (not $\mathbf{C}$) to reserve $\mathbf{C}$ for the claims matrix $\mathbf{C} \in \mathbb{R}^{N\times K}$.

CP decomposition is computed via alternating least squares (ALS):

Algorithm 1: CP-ALS

Initialise $\mathbf{A}$, $\mathbf{B}$, $\mathbf{W}$ randomly
repeat
  $\mathbf{A} \leftarrow \mathbf{X}_{(1)}(\mathbf{W} \odot \mathbf{B})(\mathbf{W}^\top\mathbf{W} * \mathbf{B}^\top\mathbf{B})^\dagger$
  $\mathbf{B} \leftarrow \mathbf{X}_{(2)}(\mathbf{W} \odot \mathbf{A})(\mathbf{W}^\top\mathbf{W} * \mathbf{A}^\top\mathbf{A})^\dagger$
  $\mathbf{W} \leftarrow \mathbf{X}_{(3)}(\mathbf{B} \odot \mathbf{A})(\mathbf{B}^\top\mathbf{B} * \mathbf{A}^\top\mathbf{A})^\dagger$
until convergence

where $\mathbf{X}_{(n)}$ is mode-n matricisation, $\odot$ is Khatri-Rao product, $*$ is Hadamard product, and $\dagger$ denotes pseudoinverse. We select rank R to achieve target explained variance $\mathrm{EV}(R) = 1 - \|\mathscr{X} - \hat{\mathscr{X}}_R\|_F^2 / \|\mathscr{X} - \bar{x}\|_F^2$; with target EV ≥ 0.90 we obtain $R = 2$ (EV = 92.45%). For robustness we also implement Tucker decomposition, $\mathscr{X} \approx \mathscr{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{W}$, where $\mathscr{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ is the core tensor and $\times_n$ the mode-n product. As the scope note explains, under the global feature-slice normalisation this factor space is degenerate for raw-scale OHLCV (Bitcoin's level dominates the leading factor), so we report no factor-based alignment leg and use $\mathbf{B}$ only as the calibration base of the positive control (Section 5.3), where its two columns are noise-corrupted to a target congruence, isometrically embedded into the ten-dimensional claims space, and recovered on those two factor axes.

C NLP Classification: Validation Detail

C.1 Semantic Taxonomy

Our taxonomy comprises $K = 10$ categories capturing core blockchain functionality (Table 5). It is author-constructed rather than adopted from a single published standard. The first two categories are the classical monetary functions of money (store of value, medium of exchange); the remaining eight are the functional use-cases that recur across the cryptocurrency-fundamentals and narrative literature (Liu et al., 2022; Shiller, 2017)—smart contracts, decentralised finance, governance, scalability, privacy, interoperability, data storage, and oracle services. We do not claim the partition is exhaustive or uniquely correct, and note alternative and finer-grained taxonomies as future work.

Table 5: Semantic Category Taxonomy

Category	**Description**
Store of Value	Digital gold, inflation hedge, wealth preservation
Medium of Exchange	Payment system, transactions, currency
Smart Contracts	Programmable contracts, automation, trustless execution
Decentralised Finance	Lending, borrowing, yield, liquidity provision
Governance	Voting, DAOs, community decision-making
Scalability	High throughput, low latency, Layer 2 solutions
Privacy	Anonymous transactions, zero-knowledge proofs
Interoperability	Cross-chain communication, bridges, multi-chain
Data Storage	Decentralised storage, file systems, permanence
Oracle Services	External data feeds, real-world information

Whitepapers are segmented into 500-word chunks ($n = 2{,}056$ across the initial 24-asset corpus; subsequently expanded and content-verified to the 43 assets analysed here). Zero-shot classification follows the entailment approach of Yin et al. (2019), constructing hypotheses of the form "This text is about [category]" for each candidate label.

C.2 Model Validation

We assess classification reliability through inter-model agreement using DeBERTa-v3 (He et al., 2021) as an alternative classifier.[4] On a random sample of 200 chunks, exact top-1 agreement is 37% (Cohen's $\kappa = 0.25$), reflecting known sensitivity of zero-shot NLI to model-specific category boundaries. Relaxed agreement—where the alternative model's top prediction appears in the primary model's top-3—reaches 68%, suggesting models capture similar semantic neighbourhoods with different decision thresholds. Recent advances using large language models (Kirtac and Germano, 2025) suggest alternative approaches for future work. Bootstrap 95% confidence intervals on aggregate category proportions (1,000 resamples) yield tight bounds—Medium of Exchange 23.1–26.2%, Data Storage 13.0–16.2%, Scalability 9.9–12.2%, Smart Contracts 9.8–11.7%, Governance 3.2–4.1%—indicating stable estimates at the corpus level despite chunk-level uncertainty.

C.3 Multi-Method Validation

To further assess robustness, we implement three independent methods with distinct inductive biases: (1) BART-MNLI (NLI-based entailment), (2) sentence embeddings using all-mpnet-base-v2 (Reimers and Gurevych, 2019) with cosine similarity to category descriptions, and (3) Ministral-3 3B, a local language model via structured JSON prompting. Table 6 reports pairwise correlations across the content-verified corpus.

Table 6: Multi-Method Classification Agreement

Method Pair	**Pearson r**	**Spearman ρ**
BART-NLI vs Embedding	0.070	0.082
BART-NLI vs LLM	0.475	0.526
Embedding vs LLM	0.395	0.333
Mean pairwise	**0.314**	**0.314**

The LLM-based classifier exhibits moderate correlation with both other methods ($r \approx 0.4$), while BART and embedding methods show weaker agreement ($r = 0.07$), suggesting distinct inductive biases. Discretised Fleiss' Kappa ($\kappa = 0.067$) indicates slight but positive inter-rater agreement above chance. Per-category agreement varies substantially (full heatmap in Appendix K): categories with clear linguistic markers show strong convergence (DeFi $\bar{r} = 0.82$, oracle $\bar{r} = 0.77$, privacy $\bar{r} = 0.71$), while abstract concepts show weaker agreement (smart_contracts $\bar{r} = 0.21$, store_of_value $\bar{r} = 0.42$). Figure 2 visualises these cross-method patterns.

[4] Model: `cross-encoder/nli-deberta-v3-small`.

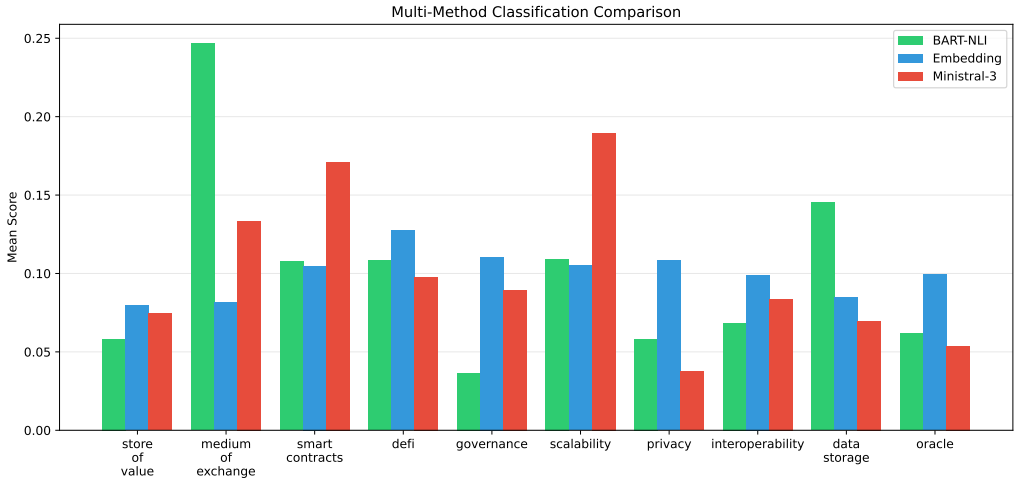

Figure 2: Multi-method classification comparison across 10 semantic categories. Bar heights represent mean classification scores for BART-NLI (primary), sentence embeddings, and Ministral-3 LLM. Categories with high cross-method agreement (DeFi, oracle, privacy) show consistent relative rankings; categories with weak agreement (smart_contracts, store_of_value) exhibit larger inter-method variance.

D Robustness Battery

This appendix reports, in full, the robustness analyses summarised in Section 5. All corroborate the bounded null.

D.1 Claims Matrix

Figure 3 displays the claims matrix heatmap, and its dominant feature is homogeneity rather than archetype. Medium of Exchange is the modal category for 36 of the 43 assets (corpus mean 24.7%), with Data Storage a consistent second (14.5%), ahead of Scalability (10.9%), DeFi (10.8%), and Smart Contracts (10.8%). The zero-shot classifier does not recover the intuitive token archetypes—a face-validity shortfall that we read as a caution on the instrument's construct validity rather than as a neutral descriptive fact (Section 6). Bitcoin's largest weight is Medium of Exchange (32.8%), not Store of Value, which ranks only sixth at 7.0%; Data Storage (18.3%) is its second category. Ethereum likewise leads on Medium of Exchange (29.2%), with Smart Contracts (14.2%) third behind Data Storage—and only the eighth-highest Smart Contracts score in the corpus, trailing BAND (16.7%), UNI (16.4%), LDO (15.3%), and MKR (14.8%). Solana and NEAR share the same Medium-of-Exchange-plus-Data-Storage backbone: Solana is led by Scalability (20.4%) and Data Storage (19.1%), NEAR by Medium of Exchange (25.7%) and Data Storage (16.3%), with Smart Contracts and Governance peripheral in both. The one category that behaves as expected is Privacy, on which Monero records the corpus maximum (18.9%, next ZEC 10.0%)—though even for Monero, Medium of Exchange (29.0%) is the larger weight. This near-uniform pull towards a generic monetary register reflects how the instrument compresses heterogeneous technical prose into a narrow band of categories. The dominant Medium-of-Exchange weight (the modal category for 36 of the 43 assets) is, however, a shared *level* effect rather than a collapse of the claims space: the column-centred claims matrix retains a participation-ratio effective rank of ≈ 6.5 (of a possible 9–10), so once that common monetary register is removed the corpus is genuinely heterogeneous, and Procrustes alignment—which optimises over rotations—is unaffected by the level. The weak claims–market correspondence is therefore not an artefact of a degenerate or near-rank-one claims matrix.

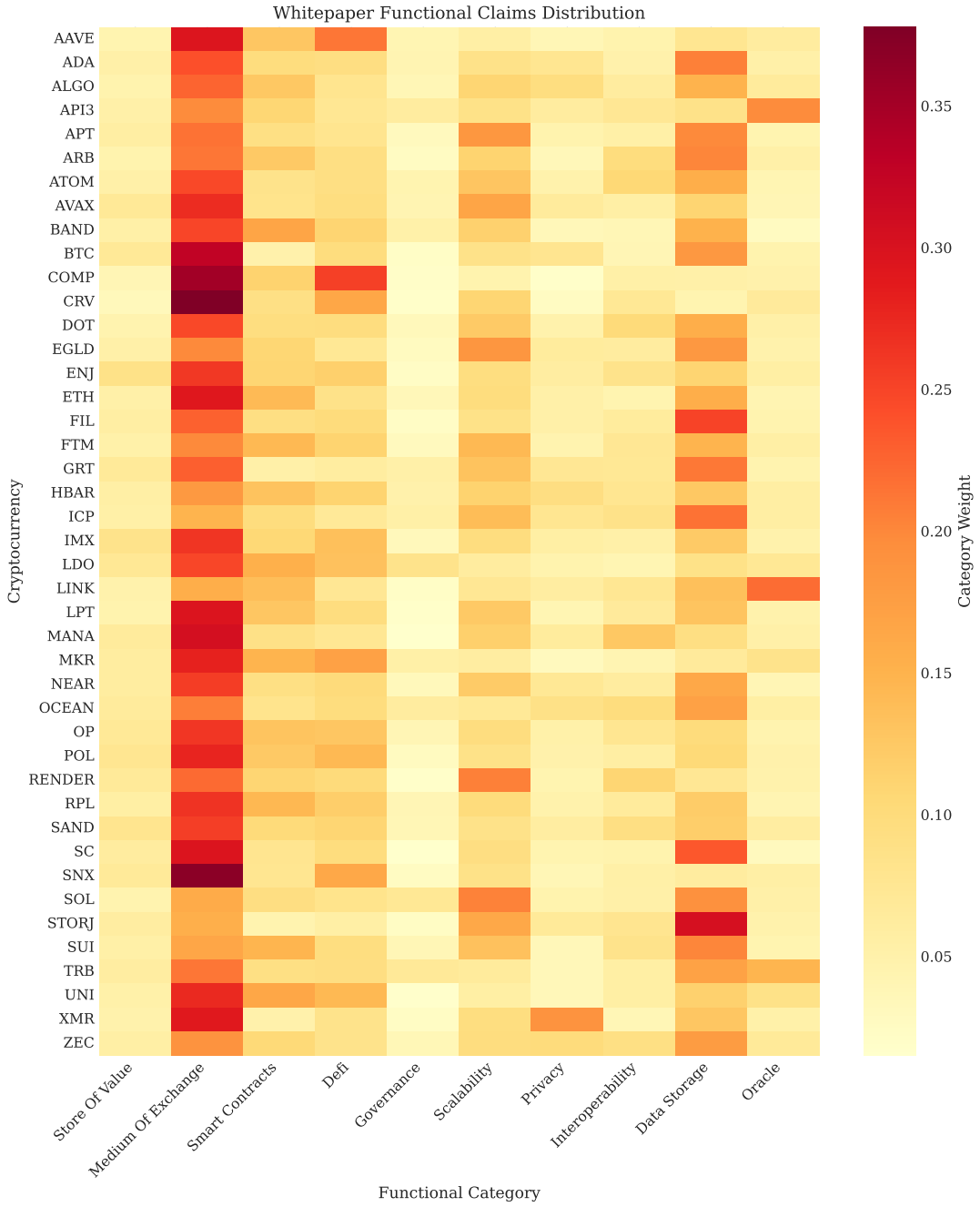

Figure 3: Claims matrix: Zero-shot classification scores across all 43 content-verified assets and 10 functional categories. Medium of Exchange is the hottest column for almost every asset, with Data Storage a consistent secondary; Monero's Privacy weight is the clearest single-category exception.

D.2 Temporal Stability

Table 7 reports alignment evolution across six rolling windows (6-month duration, 3-month stride).

Table 7: Temporal Alignment Stability Across Rolling Windows

Window	**Period**	ϕ
1	Jan–Jul 2023	0.171
2	Apr–Oct 2023	0.173
3	Jul 2023–Jan 2024	0.199
4	Oct 2023–Apr 2024	0.183
5	Jan–Jul 2024	0.113
6	Apr–Oct 2024	0.141
Mean $\pm$ SD		0.163 ± 0.028

Alignment shows moderate variation throughout the sample period, ranging from $\phi = 0.113$ (early–mid 2024) to $\phi = 0.199$ (late 2023), with all windows remaining in the weak-alignment range. The content-verified corpus (43 assets; 37 with complete coverage in every rolling window) shows comparable temporal heterogeneity, and no window approaches even moderate correspondence.

D.3 Narrative Vintage and Temporal Coverage

A natural concern is that founding-era whitepapers understate alignment because they predate the market window: a project's narrative at inception need not match the utility positioning that drives its token over 2023–2024. The textbook reading is that such vintage drift adds noise to the claims matrix relative to the contemporaneous "true" narrative and so biases observed congruence towards zero—that is, can only *attenuate* a genuine relationship, not manufacture one. We flag this as a caveat rather than a guarantee: the attenuation-only argument presumes the classifier validly captures price-relevant narrative content, which our modest inter-method agreement ($\kappa = 0.25$) calls into question—if the instrument mismeasures narrative, vintage drift is not the only force pushing ϕ down. We therefore do not treat founding-era vintage as a reason the non-detection is informative; instead we record the temporal-coverage mismatch—static, mostly pre-sample documents versus the 2023–2024 market window—as a limitation (Section 7). Directly testing whether contemporaneous, multi-period institutional narratives—documentation, governance posts, foundation updates—raise alignment requires a time-stamped corpus matched to market windows, which we leave to future work.

D.4 Feature Importance

Figure 4 shows ablation-based feature importance, detailed in Table 8.

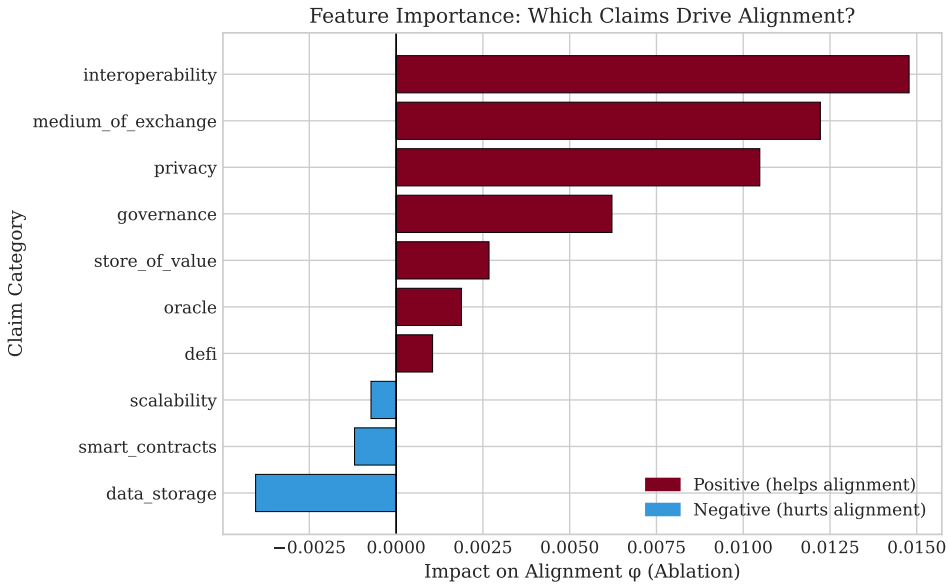

Figure 4: Feature importance via per-category ablation. All per-category contributions to alignment are small ($|\Delta\phi| \leq 0.02$), more than an order of magnitude below the 0.65 benchmark for moderate similarity; the largest deltas (interoperability, medium of exchange, privacy) are not distinguishable from noise, so no single semantic category materially rescues the weak overall correspondence.

Table 8: Feature Importance (Ablation)

Category	**Impact**
interoperability	+0.015
medium_of_exchange	+0.012
privacy	+0.010
governance	+0.006
store_of_value	+0.003
oracle	+0.002
defi	+0.001
scalability	−0.001
smart_contracts	−0.001
data_storage	−0.004

The largest per-category ablation deltas are interoperability (+0.015), medium-of-exchange, and privacy. Data-storage, smart-contracts, and scalability claims show marginally negative impact. Even the most informative category moves ϕ by less than 0.02—more than an order of magnitude below the 0.65 benchmark for moderate similarity—so no single semantic dimension rescues the weak overall correspondence.

D.5 Subsample Stability

Bootstrap resampling (100 iterations, 80% subsample) yields mean $\phi = 0.254 \pm 0.022$ with 95% CI $[0.213, 0.299]$. The point estimate is close to the full-sample result ($\phi = 0.223$) and remains firmly in the "weak" range, with the upper confidence bound well below the 0.65 threshold for moderate similarity.

D.6 Bitcoin Sensitivity

Bitcoin's exceptional position in the cross-section (the largest asset by market value and trading volume, and a multi-sigma outlier on several market statistics) raises the question of whether our results are driven by this single outlier. On the content-verified corpus ($n = 43$), Bitcoin's leave-one-out impact is negligible (-0.001), and entity-level contributions are uniformly small ($|\Delta\phi| \leq 0.022$) with no coherent niche-versus-infrastructure ordering. Bitcoin's dominant market position creates statistical leverage in the cross-section, but it does not drive the result: removing it leaves the weak, non-significant claims–market correspondence essentially unchanged.

D.7 Alternative Alignment Metrics

To ensure our results are not artefacts of the Procrustes-Tucker methodology, we supplement Tucker's ϕ with four alternative cross-space alignment measures: the RV coefficient (Robert and Escoufier, 1976), distance correlation (Székely et al., 2007), Canonical Correlation Analysis (CCA), and Partial Least Squares (PLS). These methods employ fundamentally different assumptions—RV coefficient measures configuration similarity, distance correlation detects nonlinear dependencies, CCA finds maximally correlated linear combinations, and PLS maximises covariance in latent space. If all methods converge on similar conclusions, methodological bias is unlikely. Table 9 presents results across all metrics.

Table 9: Alternative Alignment Metrics — Claims–Statistics ($n = 43$)

Comparison	**RV**	**dCor**	**CCA**	**PLS**
Claims–Statistics	0.071	0.501	0.431	0.354
	($p = 0.588$)	($p = 0.708$)	($p = 0.309$)	($p = 0.323$)

All four alternative metrics agree with the primary Tucker ϕ: the claims–statistics comparison reaches significance under *none* of them. This convergence across methodologically distinct approaches indicates that the non-detection is not an artefact of our primary Tucker ϕ measure; it does not, however, establish that the alignment is a substantive null, since all four metrics share the same low-reliability claims instrument ($\kappa = 0.25$) and none corrects for whether zero-shot classification validly captures price-relevant narrative content. The point magnitudes also caution against reading the result as a clean zero: distance correlation (0.50) and the mean canonical correlation (0.43) are moderate, not small—they simply fail significance at $n = 43$ ($p = 0.71$ and 0.31), exactly the signature of an underpowered design that cannot exclude weak alignment rather than one that has established its absence.

We compute *matched-dimension* alignment—the primary estimator reported in Section 5—by reducing the higher-dimensional matrix via SVD before computing Tucker's ϕ, avoiding any zero-padding. Reducing claims from 10D to 7D (to match statistics) yields $\phi = 0.303$ ($p = 0.365$). The pattern is unchanged: claims–statistics alignment fails significance regardless of dimension-matching strategy. This dimension-matched value is roughly $1.4\times$ larger than the zero-padded Tucker ϕ in Table 3 (the exact ratio is $1.36\times$): zero-padding averages the seven informative congruences over ten slots, deflating the coefficient by $10/7$ relative to that rotation's seven-column mean (≈ 0.319), and the separate SVD-matched estimate (0.303) lands close to that un-padded value. Both estimators nonetheless remain well below the 0.65 threshold. Because the matched-dimension ϕ is evaluated at the optimal orthogonal rotation $\mathbf{Q}^*$, 0.303 is the congruence attained at the rotation that best aligns the two spaces and cannot be raised by any other rotation. That rotation is the one maximising the Frobenius inner product $\mathrm{tr}(\mathbf{Q}^\top \mathbf{S}^\top \mathbf{C})$ over orthogonal $\mathbf{Q}$ (rotating the market statistics, $\mathbf{SQ}$, into the claims space $\mathbf{C}$), whose maximum equals

the nuclear norm $\|\mathbf{S}^\top \mathbf{C}\|_*$ by von Neumann's trace inequality; ϕ is then *evaluated* at $\mathbf{Q}^*$, and because it normalises by the column norms it is scale-invariant and does not itself equal that norm. The same estimator returns $\phi = 1$ for a market that is a perfect rotation of the claims. The low value therefore reflects weak *measured* congruence in the data, not a ceiling imposed by the estimator; as the power analysis shows, it bounds the detectable alignment rather than excluding a weak sub-threshold effect.

D.8 Scoring-Normalisation Robustness

The claims matrix is built by row-normalising the classifier's independent per-category entailment scores (Section 4). Removing that step—feeding the raw `multi_label` sigmoid scores directly into the pipeline—raises between-asset heterogeneity (the column-centred *Shannon* effective rank rises from 7.6 to 8.4, with row variance up 4.6×) and moves the matched-dimension ϕ from 0.303 to 0.336 (permutation $p \approx 0.13$). The Shannon effective rank here is the entropy convention $\exp(-\sum_i p_i \log p_i)$ on the normalised singular-value spectrum; the participation-ratio effective rank of the same row-normalised baseline (the ≈ 6.5 reported for the claims matrix earlier in this appendix) is the lower of the two summaries of that spectrum, so the two figures describe the identical baseline matrix rather than conflicting measurements. The alignment stays non-significant and far below the reliability-limited detection floor ($\phi \approx 0.44$), so the null survives this design choice.

D.9 Jackknife Stability

Leave-one-asset-out analysis for the claims–statistics alignment ($n = 43$) yields uniformly small impacts ($|\Delta\phi| \leq 0.022$): the largest positive contributors are SUI (+0.009), XMR (+0.005), and TRB (+0.005), and the largest negative are LINK (−0.022), HBAR (−0.013), and POL (−0.013). Bitcoin shows negligible impact (−0.001), confirming the null result is not driven by any single dominant asset. Monero (XMR), which anchors the Privacy dimension, was delisted from Binance in February 2024, midway through the market window; dropping it entirely leaves the result unchanged (matched-dimension ϕ falls from 0.303 to 0.289, zero-padded from 0.223 to 0.218), so the venue's mid-sample delisting does not confound the non-detection.

D.10 Controlling for Market Capitalisation

Market capitalisation may confound narrative–market relationships if larger projects have both distinctive narratives and distinctive market behaviour. We residualise all matrices on average volume (a market cap proxy) before Procrustes alignment. Controlling for market cap, claims–statistics alignment shows minimal change, remaining in the weak range (partial $\phi \approx 0.26$, versus a raw $\phi \approx 0.32$ on $n = 43$). Here ϕ is Tucker's coefficient over the seven informative dimensions without SVD reduction, so this raw baseline (≈ 0.32) is the un-reduced analogue of the matched-dimension primary estimate (0.303) rather than a separate, larger effect; the residualised and raw values are computed identically, so their comparison is internally consistent. Market cap does not drive the weak alignment result—size effects are orthogonal to the narrative-market relationship we measure.

D.11 Multiple Testing Correction

The claims–statistics comparison does not reach nominal significance under the primary Tucker ϕ, the dimension-matched estimator, or any of the four alternative metrics (RV, dCor, CCA, PLS). Because no comparison clears the uncorrected $\alpha = 0.05$ threshold, any multiple-testing correction (e.g. Bonferroni) leaves the conclusion unchanged: the non-detection is robust to multiple-testing considerations.

D.12 Bootstrap Confidence Intervals

We construct 95% CIs via percentile bootstrap ($B = 1000$ resamples). However, bootstrap resampling with replacement on small samples ($n = 43$) exhibits known pathologies when combined with Procrustes-based alignment: duplicate entities in resampled data artificially inflate ϕ by increasing effective weights on well-aligned pairs. Our bootstrap distributions show substantial upward bias (bootstrap mean exceeds point estimate by 29% for claims–statistics), with moderate right-skewness (skewness ≈ 0.45). Consequently, percentile CIs may be conservative for upper bounds but unreliable for lower bounds—the lower bound can exceed the point estimate. We therefore report these intervals for completeness while treating them as indicative rather than precise, and omit them from the headline table.

E Extended Discussion

E.1 Bitcoin's Narrative Regime

Bitcoin shows negligible leave-one-out impact (-0.001). It is nonetheless instructive that Bitcoin has transcended its whitepaper claims ("peer-to-peer electronic cash") to become a macro asset trading on "digital gold" narratives orthogonal to functional utility claims. Our alignment framework captures functional asset dynamics—the correspondence between what projects *claim to do* and how their tokens *behave*—but Bitcoin increasingly operates in a different narrative regime, dominated by macroeconomic positioning, institutional adoption, and store-of-value framing that bears little relationship to its original functional claims. On the content-verified corpus ($n = 43$), however, no token or token cluster exhibits more than negligible alignment, so this Bitcoin-specific observation is illustrative rather than evidence of a systematic cross-sectional pattern.

E.2 Feature Importance Patterns

The per-category ablation deltas are uniformly small. The largest positive are interoperability ($+0.015$), medium of exchange ($+0.012$), and privacy ($+0.010$); data storage (-0.004) and smart contracts and scalability (-0.001 each) are marginally negative. The entire spread across categories is roughly 0.02—more than an order of magnitude below the 0.65 benchmark for moderate similarity, and within the resampling variability of the overall estimate ($\phi = 0.254 \pm 0.022$). We therefore read no semantic category as substantively favoured: even the nominally most important dimension does not move ϕ enough to rescue the weak overall correspondence, and the ordering across categories is not distinguishable from noise. The ablation reinforces the non-detection rather than qualifying it.

E.3 Theoretical Implications

Our findings contribute to the growing literature on narrative economics (Shiller, 2017) by providing quantitative evidence on the limits of narrative-market coupling in cryptocurrency markets.

Narrative Dissociation Hypothesis. The weak alignment we document is consistent with what we term "narrative dissociation"—an observed weak correspondence between stated project intentions and realised market behaviour. Our limited sample size ($n = 43$) provides insufficient power to definitively distinguish weak alignment from no alignment; our results are inconsistent with strong alignment ($\phi \geq 0.70$), but narrative dissociation itself remains a working hypothesis rather than a demonstrated finding. If genuine, narrative dissociation would contrast with efficient market theory, which predicts that informative narratives are rapidly incorporated into prices. The evolving dependency structures in cryptocurrency markets (Briola and Aste, 2022) and the documented role of social media in price dynamics (Burnie et al., 2020) suggest narrative-market relationships may be more complex than our static alignment tests capture.

Bounded Rationality in Crypto Markets. The persistence of elaborate whitepaper narratives despite their apparent irrelevance to market outcomes suggests bounded rationality among market participants. Investors may allocate attention to narratives as heuristics, even when such narratives lack predictive power. This parallels findings in behavioural finance on the role of stories in investment decisions (Barberis and Thaler, 2003).

E.4 Further Limitations

Beyond the three binding limitations in Section 6, several further caveats qualify these findings:

- Whitepapers represent static documents that may not reflect current project status. Dynamic narrative analysis (social media, forum posts, governance proposals) may capture narrative evolution.
- Our functional taxonomy is author-constructed (Appendix C)—the two classical monetary functions plus eight recurring blockchain use-cases—rather than validated against an external standard; alternative partitions may reveal alignment in dimensions this one does not resolve.
- Two years of data may be insufficient to capture long-term alignment dynamics.
- Zero-shot classifiers trained on general-domain NLI corpora exhibit domain shift when applied to specialised cryptocurrency discourse (Gururangan et al., 2020). Crypto-specific terminology ("sharding," "AMM," "tokenomics") may not receive accurate treatment. We interpret this "Semantic Gap" between general-purpose NLP and crypto-native discourse as a substantive measurement challenge: off-the-shelf LLMs should not be deployed for cryptocurrency auditing or regulatory classification without domain adaptation via continued pretraining on cryptocurrency corpora.
- Single data provider (Binance) exchange prices may not represent venue-specific microstructure dynamics.

E.5 Summary of Findings

Our investigation across 43 content-verified whitepapers produced the following key findings:

1. **Claims–Market Non-Detection.** The dimension-matched congruence between claims and market statistics ($\phi = 0.303$) is weak and non-significant, well below the 0.65 threshold; the conservative zero-padded estimator ($\phi = 0.223$) is smaller still. We read this as a power-limited non-detection (Section 4.4), not as evidence of no relationship.
2. **Pipeline Liveness (Positive Control).** A positive-control simulation recovers injected cross-domain congruence almost exactly (targets $0.20/0.50/0.80$ recover to $0.234/0.501/0.806$), showing the estimator is sensitive where genuine structure exists.
3. **No Stable Entity-Level Heterogeneity.** On the content-verified corpus, leave-one-out impacts are uniformly small ($|\Delta\phi| \leq 0.022$) with no specialised-versus-infrastructure ordering; an apparent split in an earlier version of the corpus was traced to failed-download stub and wrong-document whitepapers.
4. **Temporal Dynamics.** Alignment shows moderate variation across six temporal windows ($\phi = 0.163 \pm 0.028$), ranging from $\phi = 0.113$ to $\phi = 0.199$.

5. **NLP Validation.** Inter-model agreement (BART vs DeBERTa) reaches 68% at relaxed (top-3) threshold, with bootstrap CIs indicating stable category estimates despite 37% exact agreement ($\kappa = 0.25$).

E.6 Future Work

Several extensions could strengthen this work: analysing social media content to capture narrative evolution (dynamic narratives); extending whitepaper analysis to 50+ projects (expanded corpus); fine-tuning transformer models on cryptocurrency text (alternative NLP); examining market reactions to whitepaper updates and narrative pivots (event studies); comparing alignment across blockchain ecosystems (cross-chain analysis); and extending the horizon to 5+ years as data becomes available. The cryptocurrency market remains a fascinating laboratory for studying narrative economics, market microstructure, and the relationship between information and price formation; the simple hypothesis—that projects claiming certain functionality should exhibit market behaviour consistent with those claims—is, at our sample size and instrument reliability, neither confirmed nor refuted.

F Procrustes Solution Derivation

Theorem F.1. *The orthogonal Procrustes problem*

$$\min_{\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}} \|\mathbf{AQ}-\mathbf{B}\|_F^2 \tag{7}$$

has solution $\mathbf{Q}^* = \mathbf{UV}^\top$ *where* $\mathbf{U\Sigma V}^\top = SVD(\mathbf{A}^\top\mathbf{B})$.

Proof. Expanding the objective:

$$\|\mathbf{AQ}-\mathbf{B}\|_F^2 = \mathrm{tr}[(\mathbf{AQ}-\mathbf{B})^\top(\mathbf{AQ}-\mathbf{B})] \tag{8}$$

$$= \mathrm{tr}[\mathbf{Q}^\top\mathbf{A}^\top\mathbf{AQ}] - 2\mathrm{tr}[\mathbf{Q}^\top\mathbf{A}^\top\mathbf{B}] + \mathrm{tr}[\mathbf{B}^\top\mathbf{B}] \tag{9}$$

Since $\mathbf{Q}$ is orthogonal, $\mathrm{tr}[\mathbf{Q}^\top\mathbf{A}^\top\mathbf{AQ}] = \mathrm{tr}[\mathbf{A}^\top\mathbf{A}]$ is constant. Thus we maximise:

$$\max_{\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}} \mathrm{tr}[\mathbf{Q}^\top\mathbf{A}^\top\mathbf{B}] \tag{10}$$

Let $\mathbf{A}^\top\mathbf{B} = \mathbf{U\Sigma V}^\top$. Then:

$$\mathrm{tr}[\mathbf{Q}^\top\mathbf{U\Sigma V}^\top] = \mathrm{tr}[\mathbf{V}^\top\mathbf{Q}^\top\mathbf{U\Sigma}] = \mathrm{tr}[\mathbf{Z\Sigma}] \tag{11}$$

where $\mathbf{Z} = \mathbf{V}^\top\mathbf{Q}^\top\mathbf{U}$ is orthogonal.

By von Neumann's trace inequality, $\mathrm{tr}[\mathbf{Z\Sigma}] \le \sum_i \sigma_i$ with equality when $\mathbf{Z} = \mathbf{I}$. Setting $\mathbf{Z} = \mathbf{V}^\top\mathbf{Q}^\top\mathbf{U} = \mathbf{I}$ yields $\mathbf{Q}^\top = \mathbf{VU}^\top$, and therefore $\mathbf{Q}^* = \mathbf{UV}^\top$. ■

G Tucker's Congruence Properties

Proposition G.1. *Tucker's* ϕ *has the following properties:*

1. *Bounded:* $-1 \le \phi \le 1$
2. *Scale invariant:* $\phi(c\mathbf{x},\mathbf{y}) = sign(c)\cdot\phi(\mathbf{x},\mathbf{y})$
3. *Not mean-centred (unlike Pearson correlation)*
4. $\phi = 1$ *iff* $\mathbf{x} = c\mathbf{y}$ *for* $c > 0$

H Full Asset List

The complete list of 49 cryptocurrency assets in the *market-data* (tensor) universe is: BTC, ETH, SOL, XMR, ADA, AVAX, DOT, LINK, ATOM, ALGO, FIL, ICP, AAVE, UNI, MKR, COMP, CRV, SNX, YFI, SUSHI, ENS, GRT, LDO, OP, ARB, APT, AXS, BAND, EGLD, ENJ, FTM, GALA, HBAR, IMX, LIT, LPT, MANA, NEAR, OCEAN, POL, RENDER, RPL, SAND, SC, STORJ, SUI, TRB, API3, ZEC. The claims/alignment analysis uses the 43-asset content-verified subset (Table 2); six market-listed assets are retained for tensor decomposition but lack a usable, content-verified whitepaper and are therefore excluded from the alignment tests.

I Whitepaper Corpus Details

Table 10 summarises corpus statistics for selected assets; the documents are mostly founding-era, several predating 2018, which bears on the temporal-coverage mismatch discussed in Section 6.

Table 10: Whitepaper Corpus Statistics (Selected)

Asset	Pages	Year	Type
ZEC	229	2020	Protocol Spec
STORJ	90	2018	Storage WP
NEAR	45	2020	Sharding
ICP	45	2021	Tech Overview
LINK	38	2017	Oracle WP
FIL	36	2017	Tech Report
ETH	36	2014	Original WP
ADA	32	2020	Consensus
SOL	32	2018	Original WP
MKR	21	2017	Stablecoin
XMR	20	2013	CryptoNote
+ 32 additional documents			

Documents were obtained from official project sources, academic repositories (arXiv), and GitHub. Sources include original whitepapers (BTC, ETH, SOL, AVAX), academic papers (ADA, NEAR, GRT from arXiv), protocol specifications (ZEC, LINK), DeFi protocol documentation (AAVE, COMP, MKR, UNI), storage whitepapers (FIL, STORJ, SC, AR), and technical documentation (ICP, ARB, XMR). An earlier version of this corpus contained several contaminated documents—most consequentially a GitHub-fallback retrieval that substituted the Binance Smart Chain whitepaper for Cosmos (ATOM), together with wrong-document files for ADA, NEAR, and GRT and a number of failed-download stub pages masquerading as whitepapers. These were detected by content verification (word-count thresholds and project-name provenance checks against the extracted text, not file metadata) and replaced with the correct official documents prior to the analysis reported here. The 43-asset corpus is content-verified throughout; as Section 6 discusses, the earlier contamination is itself instructive, since it manufactured a spurious cross-sectional "specialised tokens align" result that did not survive correction.

J Methodological Extensions for Future Work

Several methodological refinements could strengthen future iterations of this analysis:

Alternative Alignment Measures. The zero-padding approach for dimension-mismatched Procrustes comparison is conservative but nonstandard. Two of the alternatives are already implemented and reported in Appendix D.7 (Table 9): canonical correlation analysis (CCA) and the RV coefficient, both of which agree with the primary Tucker ϕ in returning a non-significant claims–statistics alignment (alongside distance correlation and PLS). Genuinely unaddressed extensions remain: (i) principal angles between subspaces via Grassmannian distance; (ii) HSIC for kernel-based rotation-invariant dependence; and (iii) representational similarity analysis (RSA) or Mantel tests common in cross-modal ML.

Taxonomy Validation. The ten-category taxonomy, while grounded in cryptocurrency discourse, would benefit from domain validation through expert labelling or data-driven topic discovery (e.g., BERTopic, LDA). Ablations with alternative taxonomies and finer-grained categories (L1 vs L2, DeFi subcategories, oracle networks) could reveal whether coarser groupings obscure economically salient distinctions.

Enhanced NLP Calibration. Given the modest inter-model agreement ($\kappa = 0.25$), future work should include: human adjudication on a labelled subset to calibrate zero-shot accuracy; domain-adapted few-shot prompting with chain-of-thought rationale; and sentence embedding clustering to derive data-driven categories aligned post-hoc to hypothesised domains.

Expanded Market Features. The seven aggregate statistics omit crypto-native fundamentals that may mediate narrative-market links: on-chain activity metrics (active addresses, transaction counts), token supply mechanics (inflation schedules, unlock events), total value locked (TVL) for DeFi protocols, staking yields, and developer activity (GitHub commits, contributor counts). Multi-venue data consolidation could also reduce venue-specific microstructure noise.

Dynamic Narrative Analysis. The temporal mismatch between static whitepapers (often 2017–2020) and the 2023–2024 market window may understate alignment. Rolling-window analysis with contemporaneous narrative sources (governance proposals, blog posts, Discord announcements) could test whether narrative-market coupling strengthens when narratives are temporally matched to market regimes.

K Per-Category Method Agreement

Figure 5 visualises pairwise method correlations for each semantic category. While most categories exhibit positive inter-method agreement ($r = 0.4$–0.86), smart_contracts is the weakest ($\bar{r} = 0.21$), with the embedding–LLM pair essentially uncorrelated ($r = -0.03$), suggesting this category's linguistic markers are interpreted differently across model architectures. Categories with clearer linguistic anchors (DeFi, oracle, privacy) show strongest convergence.

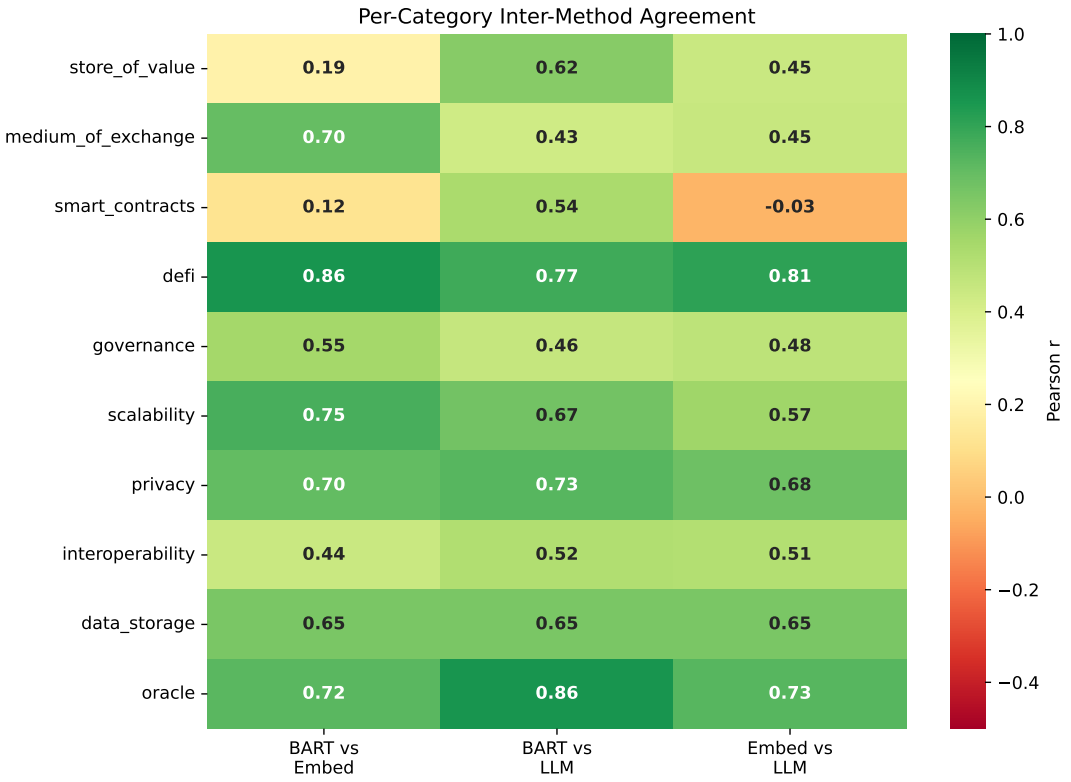

Figure 5: Per-category Pearson correlations between three classification methods: BART-NLI vs Embedding, BART-NLI vs LLM, and Embedding vs LLM. Most categories show moderate positive agreement (green), but smart_contracts exhibits negative Embedding–LLM correlation (red), indicating model-specific interpretation of this technically ambiguous category.