

The Confidence Premium

Peer-relative CEO confidence in earnings-call language as a cross-sectional equity signal

Universe 2,500 largest US-listed companies, 3,504 names, survivorship-proof **Sample** 2011-07-13 to 2026-07-10, 3,913 evaluation dates **Horizon** 63 trading days

Signal tx_ceo_confidence_vs_peers **Model** Binomial Marks v1 **Classification** Professional / institutional use

ABSTRACT

Each quarter, several thousand executives narrate their results on earnings calls. The Binomial Marks model reads every transcript and scores how confident each management team sounds. We then rank that confidence against the peers reporting the same fiscal quarter. This note asks whether that peer-relative confidence predicts subsequent returns across the 2,500 largest US-listed companies over 2011 to 2026. It does. The effect is small but statistically robust: a 63-day rank information coefficient of +0.025 (Newey-West t of 3.24), positive at every horizon from two weeks to six months and in every subperiod examined. It survives a multiple-testing (deflated-Sharpe) penalty, and, unusually for a language signal, a nine-factor style-spanning regression leaves approximately 4.9 per cent per annum of residual alpha that the established style factors do not explain. As a dollar-neutral book the net Sharpe ratio is a modest 0.4 to 0.6 after realistic costs; the economic value increases under a light net-long tilt, at which point the Sharpe ratio approaches 1.2. The principal risk is a rotation back towards value.

+0.025

Rank IC, 63d

3.24

Newey-West t

65.6%

Daily hit rate

+4.93%

Residual alpha p.a.

0.69

Deflated Sharpe

1.22

Sharpe, long tilt

1 Signal source: the Binomial Marks model

The raw material of this study is not a figure that a company reports but the language its management uses to describe that figure. The reported numbers are priced within milliseconds. The tone of the call, whether a chief executive sounds confident and specific or defensive and evasive, is considerably harder to measure and, on the evidence below, is priced more slowly.

Binomial Marks v1 is the specialist language model that produces the confidence reading. It is a 400-million-parameter long-context transformer encoder. It takes a full earnings-call transcript, up to a 16,384-token context that covers the 99th percentile of calls, together with a deterministic meta-data prefix that carries sector, country, ticker and fiscal quarter, and returns 23 structured values per call. This sector-aware conditioning is the model's main source of edge: an identical sentence about margin pressure is scored as a sharper negative in software, where margins are typically stable, than in energy, where margin volatility is structurally higher.

The 23 outputs comprise ten topic-direction scores, each signed on $[-2, +2]$ with a binary mention flag, and three tone scores, each a regression on $[1, 5]$. The tone dimension used in this study is `mgmt_confidence`.

Two properties of the model matter for what follows. The first is determinism: identical inputs always produce identical outputs, so a backtest is reproducible without freezing external responses. The second is a stated limitation: tone levels are partially mode-collapsed, with confidence clustering around 4 to 5, so the model captures relative ordering

Table 1. Tone dimensions, scale 1 to 5

	1 denotes	5 denotes
<code>mgmt_confidence</code>	uncertain, hedged	direct commitments
<code>mgmt_defensiveness</code>	open, direct	evasive, deflecting
<code>analyst_scepticism</code>	soft, congratulatory	visible scepticism

well whilst absolute levels drift across sectors and over time. That limitation is the reason the factor is constructed as a peer-relative rank rather than as a raw score.

Table 2. Validation against a frontier-LLM panel

Metric	Marks	Frontier
Topic-direction Spearman	0.691	0.838
Tone Spearman	0.62	0.61
Mention macro-F1	0.909	0.95
Confidence (tone) Spearman	0.66	–

Labels are distilled from a panel of frontier reasoning models. Marks reproduces approximately 82 per cent of the residual agreement the frontier models achieve with one another (0.691 against 0.838), whilst running deterministically in around 50 ms on a CPU at negligible marginal cost. Of the three tone dimensions, confidence is the cleanest, since its language is the most explicit.

2 Signal construction

The Marks model supplies only a raw confidence reading for each call; the peer ranking is ours. That step is the decisive one: we grade each executive against the peers reporting the same fiscal quarter. Tone varies by industry and by point in the cycle, so an absolute confidence score is not comparable

across companies. Ranking within the peer set removes that common variation and isolates the idiosyncratic component, namely how confident a given management team sounds relative to everyone else describing the same macroeconomic quarter.

$$s_{i,t} = \text{rank}_{j \in P(i,t)}(c_j) - \frac{1}{2}(N_t + 1), \rightarrow \mathcal{N}(0, 1).$$

Here c_j is the Marks confidence reading for company j ; $P(i, t)$ is the peer set reporting the same fiscal quarter within the sector; and N_t is the size of that peer set. The signal is then standardised to zero mean and unit variance. A score of +1 corresponds to roughly one rank above the median peer for the quarter. Only the relative dimension carries predictive power: a simple time-series average of a company's own confidence has no value for trading.

Coverage. The Marks model reaches back to 2006 and covers approximately 80 per cent of large-cap names, sufficient to construct a genuinely diversified long/short book. The evaluation sample used here spans 3,504 names over 3,913 daily cross-sections from 2011-07-13 to 2026-07-10, on a point-in-time top-2,500-by-market-capitalisation universe that is survivorship-proof.

3 Methodology

The core test is the information coefficient. On each date, companies are ranked by the confidence signal and again by the return they subsequently earn over the following h days; the agreement between the two orderings is measured and then averaged across all dates.

$$\text{IC}_t = \rho_{\text{Sp}}(s_{\cdot,t}, r_{\cdot,t \rightarrow t+h}), \quad \overline{\text{IC}} = \frac{1}{T} \sum_{t=1}^T \text{IC}_t.$$

An IC of +1 is a perfect ordering and 0 is no information. In equities a genuine edge is small and steady: even strong quantitative signals rarely exceed 0.05. Because the forecast looks h days ahead on every date, consecutive readings overlap and are not independent, which inflates a naive t -statistic. This is corrected with Newey-West standard errors (Bartlett kernel, lag $h - 1$), and the horizon family is adjusted with the Benjamini-Hochberg procedure.

$$t_{\text{NW}} = \frac{\overline{\text{IC}}}{\widehat{\text{se}}_{\text{NW}}(\overline{\text{IC}})}, \quad \text{Spread}_t = \bar{r}_{Q10,t} - \bar{r}_{Q1,t}.$$

Two further tools appear later. A Fama-MacBeth regression estimates a cross-sectional slope on the signal each date and reports the time-series mean, an independent check on the IC. The Deflated Sharpe Ratio asks whether an observed result exceeds the best outcome expected from having examined several candidate signals, adjusting for skewness $\hat{\gamma}_3$ and kurtosis $\hat{\gamma}_4$ of the return stream.

$$\widehat{\text{DSR}} = \Phi \left(\frac{(\widehat{\text{SR}} - \text{SR}_0) \sqrt{T-1}}{\sqrt{1 - \hat{\gamma}_3 \widehat{\text{SR}} + \frac{\hat{\gamma}_4 - 1}{4} \widehat{\text{SR}}^2}} \right).$$

4 Empirical results

The strength of the relationship is stable across horizons. Averaged over approximately 3,800 trading days, the 63-day IC is +0.025. It remains positive at every horizon from two

weeks to six months, and strengthens slightly as the horizon lengthens, consistent with tone being priced slowly.

The relationship is monotone when names are sorted into ten confidence buckets: average forward return rises almost step by step from the least confident decile to the most confident. An independent Fama-MacBeth regression agrees, placing the premium for moving up the confidence ranking at approximately +1.76 per cent per annum ($t_{\text{NW}} = 3.77$). The predictive strength is persistent rather than the product of a single favourable period.

5 Robustness

Any researcher who examines enough ideas will eventually find one that appears convincing by chance. Guarding against that possibility is the central discipline. Two checks are applied: stability across time, and an explicit penalty for the number of trials.

Stability across subperiods and years

A genuine effect should appear in every slice of the history. This one does. The signal is positive in both halves of the sample, in all three thirds, and in 12 of the 16 individual calendar years.

The penalty for multiple trials

Suppose that a dozen variations had been examined quietly and only the best reported. To price that risk, a null distribution is built with a 500-path block bootstrap (21-day blocks), and the observed result is located within it.

The verdict is a Deflated Sharpe of 0.69 and a probabilistic Sharpe versus zero of 0.98. The effect comfortably survives the multiple-testing penalty. For calibration, the best of the other candidate signals screened for this series collapsed to 0.17 under the identical test, so survival is not automatic.

The sector footprint is uneven from year to year (Exhibit 7). It is strongest in Real Estate, Communication Services and Consumer Defensive, and weakest in Basic Materials, where a small number of extreme years dominate. This argues for sector-neutral construction in production.

6 Real alpha, or a growth tilt in disguise

The objection a professional raises immediately is that confident chief executives may simply run faster-growing companies, in which case this is merely the established growth factor, already available cheaply in an exchange-traded fund. To test this, the strategy's returns are regressed on nine standard style factors and the intercept, or alpha, the return that none of the known factors explains, is examined.

The loadings are real. Confident-management firms genuinely tilt towards growth ($\beta = +2.15$) and away from value ($\beta = -1.67$). After removing all nine styles, however, approximately +4.9 per cent per annum survives as residual alpha the style factors do not explain, though at $t = 1.69$ it is suggestive rather than decisive in isolation.

The contrast with a sibling signal is what settles the question. A near-twin from the same Marks engine that scores management's guidance tone predicts returns comparably well on paper, yet once the style factors are removed essentially nothing remains. Same model, two signals, and only the confidence dimension carries something the market has not already priced.

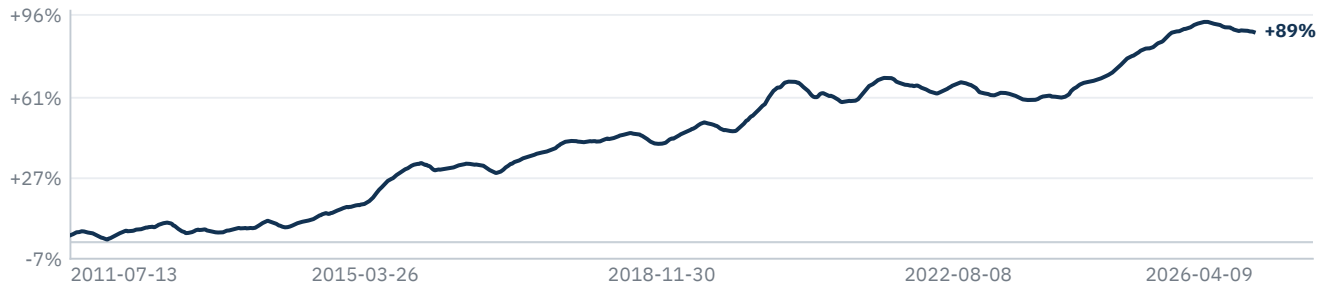
Table 3. Information coefficient by forecast horizon

Horizon	Mean IC	$\sigma(\text{IC})$	t naive	t Newey-West	p BH	Hit rate	n dates
14 days	+0.021	0.082	15.78	5.60	0.000	63.9%	3,869
21 days	+0.024	0.083	17.61	5.21	0.000	65.3%	3,861
42 days	+0.025	0.082	18.70	3.91	0.000	65.7%	3,841
63 days (primary)	+0.025	0.080	19.02	3.24	0.001	65.6%	3,820
126 days	+0.027	0.077	21.05	2.66	0.008	63.2%	3,753

The naive t is upward-biased under overlapping horizons; the Newey-West column is the reliable one. Every horizon clears conventional significance after the Benjamini-Hochberg adjustment.

Cumulative long/short spread

Top-minus-bottom confidence decile, cumulative %, de-overlapped, gross of costs

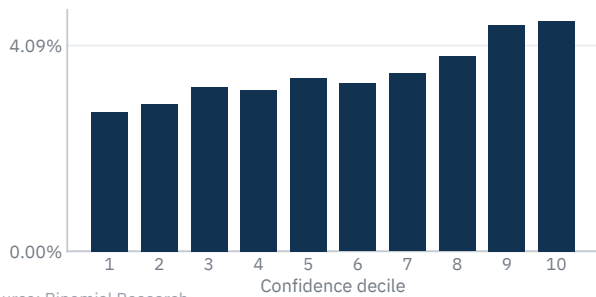


Source: Binomial Research.

Exhibit 1. Forward-return spread of the top-minus-bottom confidence decile at the 63-day horizon, de-overlapped and gross of costs. The steady climb over fifteen years is the central result.

Return by decile

Avg forward return, %



Source: Binomial Research.

Exhibit 3. Average forward return by decile, pooled across evaluation dates at the primary horizon.

Table 4. Subperiod stability

Period	Span	Mean IC	t NW	n
H1	2011–2018	+0.026	2.78	1,910
H2	2018–2026	+0.023	1.95	1,910
T1	2011–2016	+0.028	2.66	1,273
T2	2016–2021	+0.022	1.38	1,273
T3	2021–2026	+0.024	1.97	1,274

That call language predicts returns is broadly true; that call language is free, undiscovered alpha is true specifically for confidence.

An orthogonality scan against the full catalogue of known signals confirms the picture: the highest absolute cross-sectional correlation is 0.10, against an up-capture measure, far below the 0.7 threshold that would flag a repackaged factor. On this evidence the confidence signal stands on its own.

7 Portfolio implementation

Predicting returns and realising them are different matters. Realising them requires a live portfolio, trading costs and tol-

Table 5. Year-by-year mean IC and hit rate

Yr	IC	Hit	Yr	IC	Hit
2011	−0.033	31%	2019	+0.040	65%
2012	+0.012	69%	2020	+0.001	54%
2013	+0.028	82%	2021	+0.022	60%
2014	+0.022	74%	2022	−0.005	39%
2015	+0.075	86%	2023	+0.000	41%
2016	−0.007	49%	2024	+0.079	95%
2017	+0.062	89%	2025	+0.046	75%
2018	+0.010	62%	2026	−0.010	61%

2026 covers a partial year (18 evaluation dates). The four negative years, 2011, 2016, 2022 and 2026, coincide with value-led or highly defensive markets, consistent with the factor's growth tilt discussed in Section 6.

Table 6. Bootstrap and deflation summary

Statistic	Value
Replications / block	500 / 21d
Observed IC	0.025
Bootstrap IC (p5, p50, p95)	0.003, 0.013, 0.024
Bootstrap p -value (one-sided)	0.042
Deflated Sharpe (12 trials)	0.689
Probabilistic Sharpe vs 0	0.985

Table 7. Style-factor loadings, largest $|\beta|$

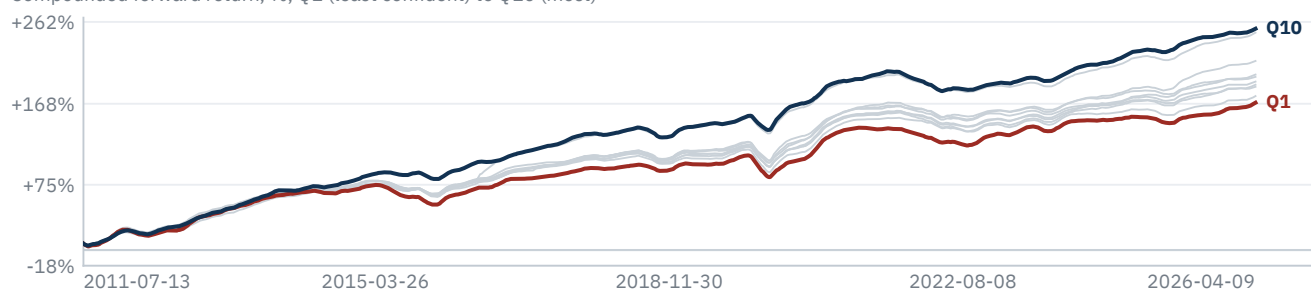
Style factor	Loading β	t NW
Growth	+2.155	4.44
Value	−1.673	−3.87
Size	+0.757	1.67
Leverage	+0.502	1.06
Momentum	+0.253	1.30
Volatility	−0.242	−1.87
Quality	+0.191	0.41
Beta	+0.178	1.66
Residual alpha (p.a.)	+4.93%	1.69

R^2 on factor returns = 0.44; observations = 1,835. A high R^2 with an insignificant alpha would indicate a repackaged known factor, which is not the case here.

erance for drawdowns. The simulation reflects exactly that: each month the most confident companies are bought and the least confident sold, under a fully disclosed cost model of approximately 26 basis points round-trip plus 1 per cent per annum borrow on shorts and square-root market impact.

Cumulative return by confidence decile

Compounded forward return, %, Q1 (least confident) to Q10 (most)



Source: Binomial Research.

Exhibit 2. Cumulative forward return by decile. Q10 (navy) sits above and Q1 (red) below, with interior deciles in grey, indicating a genuine ordering rather than luck in the extremes.

Rolling information coefficient

21-day EWMA of the daily cross-sectional rank IC, 63-day horizon



Source: Binomial Research.

Exhibit 4. Rolling information coefficient. Persistence of sign matters more than level.

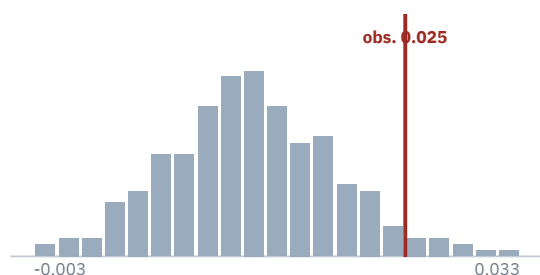
Table 8. Portfolio grid, tranche by net exposure

Configuration	Tranche	Net	CAGR	Sharpe	Sortino	Max DD	Win%	Turn%	Names L/S	IR vs SPY
Market-neutral	top/bottom 20%	0.0	+1.3%	0.44	0.64	-5.2%	62%	50	256 / 257	-0.80
Market-neutral	deciles	0.0	+1.5%	0.40	0.59	-6.9%	61%	56	128 / 129	-0.77
Market-neutral	top/bottom 5%	0.0	+1.8%	0.42	0.63	-8.2%	59%	62	65 / 65	-0.76
Light long tilt	deciles	+0.3	+6.8%	1.22	2.25	-10.2%	70%	56	129 / 129	-0.58
Long-only top decile	deciles	+1.0	+18.2%	1.11	1.88	-28.5%	69%	54	128 / 0	+0.89

Monthly rebalance, top-2,500 by market capitalisation. Sharpe on month-end mark-to-market NAV ($\times \sqrt{12}$). Net denotes target net market exposure. The net-spread and deflation figures in Section 5 use a lighter 5 bps one-way assumption (0.22 per cent annual cost drag); the grid above uses the fuller 26 bps round-trip plus borrow and impact model.

Bootstrap null vs observed

Block-bootstrap IC distribution



Source: Binomial Research.

Exhibit 5. Block-bootstrap null from resampled returns; the red line is the observed IC, in the right tail. One-sided $p = 0.042$.

Volatility is measured month-end to month-end.

Three conclusions follow. First, fully hedged, the effect is real but small: dollar-neutral, the net Sharpe ratio is near 0.4 after costs, dependable rather than dramatic, and concentrating into the extreme 5 per cent adds trading cost without improving the ratio. Second, the economic value ap-

pears with a modest amount of market exposure: a light long tilt raises the Sharpe ratio to 1.22, because the book then collects the confidence effect and the equity risk premium, which diversify one another. Third, as a long-only tilt the strategy outperforms the market, compounding at 18.2 per cent per annum with a positive information ratio against the S&P (+0.89), at the cost of a deeper drawdown of -28.5 per cent, since at that point it is largely a market position with a confidence tilt applied.

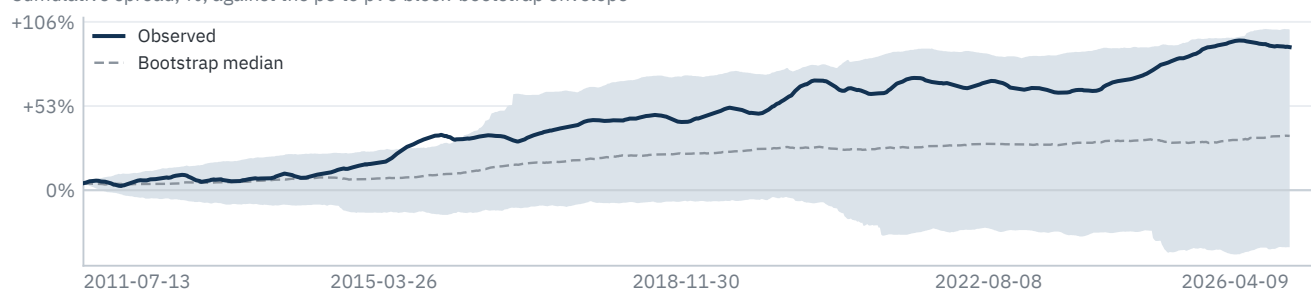
8 Thesis, risks and provenance

Thesis. Management tone is priced slowly. Ranked within sector and fiscal quarter, the management teams that sound most confident relative to their peers tend to outperform over the following quarter. The signal is small but unusually clean; it survives deflation and style-spanning and is not merely a repackaged growth tilt.

Principal risk. A rotation back towards value. Given the growth-long and value-short loadings ($\beta_{\text{growth}} = +2.15$, $\beta_{\text{value}} = -1.67$), the strategy will underperform in value-led markets, as observed in the negative IC years of 2011, 2016 and 2022. Production use argues for sector-neutral, style-hedged construction.

Observed spread versus bootstrap band

Cumulative spread, %, against the p5 to p95 block-bootstrap envelope



Source: Binomial Research.

Exhibit 6. Observed cumulative spread (solid navy) against the p5 to p95 bootstrap envelope, with the bootstrap median dashed. The observed curve tracks the upper edge of the band throughout.

Sector by year mean information coefficient

Cross-sectional rank IC within sector, averaged per calendar year. Blue positive, red negative.

	'11	'12	'13	'14	'15	'16	'17	'18	'19	'20	'21	'22	'23	'24	'25	'26
Basic Materials	-.20	-.06	-.04	.04	.09	-.02	-.05	-.01	.06	-.03	.13	.03	-.07	.11	.31	.00
Comm. Services	.10	.20	.22	.05	.02	.09	.04	.08	-.04	.07	-.07	.09	.06	.01	.01	-.02
Consumer Cyclical	.03	-.03	-.06	.04	.04	.03	.13	-.04	.07	.06	-.01	.03	.06	.06	-.03	.05
Consumer Defensive	.08	.09	.03	-.02	-.02	-.07	.11	.03	.05	-.11	.01	-.14	.06	.10	.09	.03
Energy	.06	.04	.05	-.02	.04	-.08	.03	-.01	.03	.05	.08	.05	-.09	.15	-.07	.04
Financials	-.11	.06	.07	.05	.05	-.03	.02	-.03	.06	.02	.00	.00	-.04	.02	.03	-.01
Healthcare	-.05	.05	.16	-.03	.05	.05	.08	.11	.00	-.01	.01	.04	-.03	.07	.05	-.02
Industrials	-.02	-.01	.04	.04	.10	-.05	.09	-.04	.05	-.06	.06	-.01	.04	.11	.15	-.04
Real Estate	.12	.09	.04	.22	.23	.06	.14	-.04	.10	-.07	.05	.05	.03	-.01	.01	.04
Technology	-.06	.08	-.01	-.01	.11	.04	.13	.07	.03	.02	-.01	-.03	.01	.08	-.05	-.05
Utilities	.05	.05	.11	.02	.09	.02	.03	-.03	.05	.08	.09	-.03	-.07	.07	-.07	.01

Exhibit 7. Sector by year mean information coefficient, shown as a heatmap. Blue cells denote a positive within-sector IC for the year, red cells a negative one; deeper colour indicates a larger magnitude.

Table 9. Definition and provenance

Field	Value
Model	Binomial Marks v1
Signal	tx_ceo_confidence_vs_peers
Universe	pit_top_n, top 2,500
Start	2011-07-13
Primary horizon	63 days
Names / dates	3,504 / 3,913

Important disclosures. This document is provided for informational purposes only and is intended for institutional and professional investors. It does not constitute investment advice, a research recommendation, or an offer or solicitation to buy or sell any security or to adopt any investment strategy. The analysis is based on hypothetical, backtested performance. Backtested results are simulated, do not represent the results of actual trading, and have inherent limitations; no representation is made that any account will, or is likely to, achieve results similar to those shown. Past performance is not a reliable indicator of future results. Investing involves risk, including the possible loss of principal. Information contained herein has been obtained from sources believed to be reliable, but its accuracy and completeness are not guaranteed. Views expressed are as of the date of publication and are subject to change without notice. © 2026 Binomial Technologies. All rights reserved.