Output-Format Granularity as a Reference-Free Diagnostic for LLM Annotation: A Case Study on Earnings-Call Polarity

Anonymous ACL submission

Abstract

LLMs are increasingly deployed as zero-shot annotators in domains where sentence-level gold labels are unavailable. This creates a validation problem: downstream predictive validity can show that labels carry signal, but it cannot show whether the annotation scheme is stable under distribution shift, nor diagnose which design choices break when it is not. We propose a reference-free diagnostic that holds the corpus, construct, and downstream task fixed while varying the granularity of the LLM output format, from lexicon counts through binary and categorical tagging to continuous Likert scoring. The diagnostic object is the pattern of results across schemes rather than any single coefficient. We instantiate the protocol on forward-looking polarity in earnings-call transcripts, validated against post-earnings-announcement drift. On a US panel of 75 firms (1,789 firm-quarters), LLM categorical tagging of negative forward-looking tone after positive earnings surprises predicts lower residualized returns over trading days 6 to 30 ($t = -2.50$). The same categorical scheme reverses sign on a 166-firm Asian panel ($t = +2.37$), with the reversal concentrated in 23 HKEX firms with mainland-China exposure listed via Cayman holding structures ($t = +2.86$). The reversal is not a sector effect, since the closest US technology and communication-services subsample retains the expected sign ($t = -1.56$). Critically, on the same firms the failure attenuates when the output format is changed from categorical bins to averaged ordinal Likert scoring ($t = +1.13$). We find no significant difference in within-sentence concession-marker frequency between US and PRC-via-Cayman sentences, so we do not claim a simple cultural-register mechanism. The methodological takeaway is that LLM annotation pipelines should be validated across output formats, not just across prompts or models, since granularity itself can determine whether a construct transfers.

1 Introduction

LLMs are now used as zero-shot annotators in research settings where gold labels are expensive, unavailable, or conceptually unstable (Gilardi et al., 2023; Zhu et al., 2023; Törnberg, 2023). This is attractive in domains such as finance, law, medicine, and political communication, where researchers often need labels for nuanced textual constructs rather than standard benchmark categories. It also raises a validation problem. If no ground-truth labels exist, how can one tell whether an LLM annotation pipeline measures the intended construct?

A common answer is downstream validation. Labels are taken to be credible if they predict an external outcome of interest. This is useful (and common practice) but incomplete. Downstream prediction can show that an annotation pipeline extracts something relevant, but it cannot tell us whether the construct is robust under distribution shift, whether the model is overfitting a domain-specific proxy, or whether the output format itself is producing a brittle decision rule. Inter-model agreement helps, but models can agree for the wrong reason.

We propose a reference-free *granularity test*. Rather than evaluating a single annotator, the researcher annotates the same corpus with a family of schemes that vary how much structure the LLM output is forced into, from simple word counts at one end to free-form continuous scores at the other. Each scheme is then evaluated against the same downstream task, and the diagnostic object is the pattern of results across schemes. If adjacent schemes behave similarly, the underlying construct is more credible. If some schemes break under distribution shift while others do not, the divergence localizes the failure to a specific design choice, such as forcing the model into discrete categories versus allowing continuous scores.

We instantiate the protocol on forward-looking polarity in earnings-call transcripts. This is a use-

ful testbed for three reasons. First, earnings calls contain many forward-looking statements, but no sentence-level gold label exists for whether a sentence is future-oriented, negative, vague, or hedged. Second, the finance literature provides a clean downstream validation target in post-earnings-announcement drift (PEAD): after a firm reports earnings that beat expectations, its stock price tends to continue drifting in the same direction over the following weeks (Ball and Brown, 1968; Bernard and Thomas, 1989; Livnat and Mendenhall, 2006; Meursault et al., 2022). The size of the earnings beat, measured by standardized unexpected earnings (SUE), is the standard control variable. Third, prior work shows that financial-text sentiment behaves differently outside the US, which makes cross-market transfer a realistic stress test (Pinto and Siano, 2025).

The empirical setting has two corpora. The US corpus contains 1,789 firm-quarter observations from 75 large-cap firms. The Asian corpus contains 2,560 firm-quarter observations from 166 firms listed on the Hong Kong (HKEX) and Singapore (SGX) exchanges, including a 23-firm subgroup of HKEX-listed companies with mainland-China operating exposure that incorporate offshore through Cayman or similar holding structures, henceforth "PRC-via-Cayman". We annotate all sentences in both corpora with six schemes that span lexicon counts, lexicon-flagged sentences, binary LLM classification, categorical LLM tagging, structured LLM field extraction combined with deterministic polarity rules, and continuous 1–5 LLM scoring. Figure 1 summarizes the diagnostic protocol.

The paper reports three main results. First, the US sample shows convergent downstream validity. The LLM extraction-plus-rules scheme produces the strongest PEAD signal: after positive earnings surprises, a higher share of negative forward-looking sentences predicts lower residualized returns over trading days 6 through 30 ($t = -2.50$). A next-quarter fundamentals test points the same way. The share of negative forward-looking sentences from the categorical-tagging scheme predicts next-quarter SUE with $t = -3.06$, and the continuous-Likert confidence score does so with $t = -2.68$.

Second, the same extraction-plus-rules measure fails in Asia. The coefficient reverses sign on the full Asian panel ($t = +2.37$), and the reversal concentrates in the 23 PRC-via-Cayman firms ($t = +2.86$); removing those firms leaves a weak, insignificant result. The reversal is not driven by sector composition. The closest US analogue, the technology and communication-services subsample, retains the expected negative sign ($t = -1.56$).

Third, the failure attenuates as output-format granularity increases. On the same PRC-via-Cayman firms, the extraction-plus-rules scheme gives the strongest wrong-sign result ($t = +2.86$), categorical tagging without extraction is weaker ($t = +1.47$), and averaged ordinal Likert scoring is weakest ($t = +1.13$). This is the central diagnostic finding. The failure is not simply that LLMs fail on Asian earnings calls. It is sharper than that. Forcing the model to make a categorical sentence-level polarity decision fails more strongly than asking the same model for an ordinal 1–5 polarity score on the same sentences, averaged across the call.

We initially expected this pattern to be explained by within-sentence concession structures such as "challenges, but opportunities," which can plausibly confuse a categorical polarity tagger. Qualitative examples remain useful for illustration, but aggregate tests do not support a simple frequency-based mechanism. PRC-via-Cayman sentences do not contain concession markers at significantly higher rates than US sentences. We therefore frame the contribution conservatively. The paper localizes a transfer failure to output-format granularity and leaves the deeper discourse-level mechanism open for future work.

2 Related Work

2.1 LLMs as annotators without gold labels

LLM-as-annotator workflows often rely on model agreement, prompt stability, or downstream predictive validity (Gilardi et al., 2023; Törnberg, 2023; Zhu et al., 2023; Chiang et al., 2023). These checks answer related but different questions. Model agreement asks whether multiple systems make similar decisions. Prompt stability asks whether small wording changes alter the output. Downstream validity asks whether labels predict an external outcome. None of them directly isolates the role of output format.

Our protocol complements these checks by treating annotation design as an experimental factor. The same corpus, construct, model family, and downstream task are held fixed while the granularity of the annotation output changes. The goal is not to prove that one scheme is universally best, but to reveal whether a result is stable across rea-

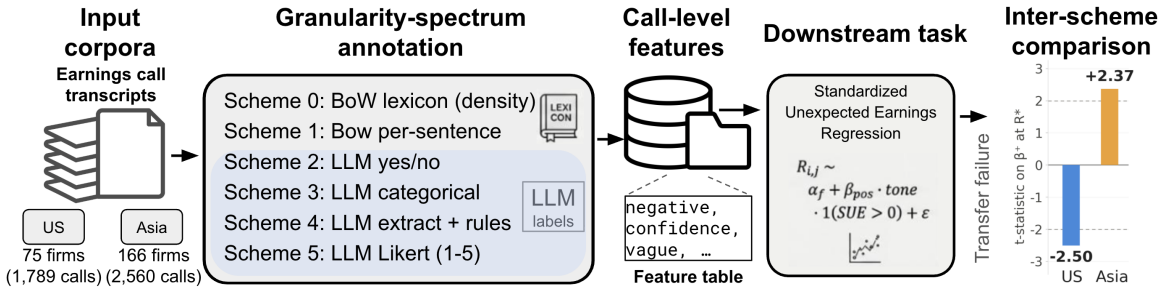

Figure 1: Granularity-spectrum validation protocol. The same corpora are annotated with six schemes, collapsed to call-level features, and evaluated on a common downstream task.

sonable formulations of the same latent construct, and to diagnose which formulation breaks under transfer.

2.2 Financial text, sentiment, and PEAD

Financial-text research has long used dictionaries and domain-specific models to measure disclosure tone (Tetlock, 2007; Loughran and Mcdonald, 2011). Loughran and McDonald introduced a finance-specific sentiment dictionary for corporate disclosures, and FinBERT(Araci, 2019) and related transformer models improved contextual financial sentiment modelling. More recent work shows that LLM methods can extract information from earnings disclosures beyond conventional dictionary methods, and PEAD.txt shows that text-derived earnings-announcement information can predict post-announcement drift even when classical PEAD has weakened (Siano, 2025; Meursault et al., 2022).

Our paper is not primarily a new trading-signal paper. The downstream finance task serves as a reference-free validation environment for annotation. The finance contribution is narrower. The LLM tone signal is predictive in US earnings calls, but its transfer to Asian English-language earnings calls is not stable, and the failure depends on output format.

2.3 Cross-cultural financial disclosure and discourse structure

Recent international disclosure work shows that sentiment-return relations are weaker outside the US and that cultural distance matters (Pinto and Siano, 2025). This paper extends that line in two ways. First, it tests LLM annotation rather than dictionary sentiment. Second, it studies a specific failure mode rather than only documenting average attenuation.

The qualitative examples also connect to discourse-structure work on concession and polarity reversal (Taboada and Mann, 2006; Heerschop et al., 2011; Zhou et al., 2011). Sentences containing contrastive relations can defeat simple polarity classifiers because the sentiment-bearing clause and the discourse nucleus need not align. However, our aggregate tests find no significant difference in concession-marker frequency across subgroups, so we treat the discourse literature as motivation rather than as a completed explanation of the empirical failure.

3 Data

3.1 US panel

The US sample contains 75 large-cap firms with continuous coverage in the Center for Research in Security Prices (CRSP) database and matched quarterly fundamentals from the S&P Compustat North America database, both accessed via Wharton Research Data Services (WRDS). Earnings-call transcripts come from S&P Capital IQ and cover 2018–2025. After merging transcripts, quarterly fundamentals, earnings-surprise measures, and residualized daily returns, the matched panel contains 1,789 firm-quarter observations for the main drift test and 1,715 observations for the next-call SUE test.

Earnings surprise is constructed using the Livnat–Mendenhall seasonal-difference approach (Livnat and Mendenhall, 2006) and winsorized at the 1st and 99th percentiles. Daily returns are residualized using firm-level factor regressions. The main outcome is the residualized buy-and-hold return from trading day 6 to trading day 30 after the call, denoted R^* (fully defined later).

3.2 Asian sample

The Asian sample contains 166 HKEX and SGX firms with English-language earnings-call transcripts from 2018–2025, and the matched panel

contains 2,560 firm-quarter observations. Fundamentals come from Compustat Global. Because the same US earnings variables are not consistently available in Global, Asia SUE is based on quarterly income before extraordinary items. Returns are residualized using regional factor returns.

We identify 23 PRC-via-Cayman firms, defined as companies with mainland Chinese operating exposure that list in Hong Kong through Cayman or similar offshore holding structures. This subgroup includes large China-internet and platform firms. The flag is used only for subgroup analysis after observing the full Asia failure.

4 Annotation Schemes

Table 1 defines the six annotation schemes, which span a granularity spectrum by design along two stacked dimensions: what the model is asked to output per sentence, and how those sentence-level outputs are aggregated to call-level features.

Sentence-level output. Schemes 0–1 are lexicon baselines that require no model. Scheme 0 counts forward-looking words in the full call; Scheme 1 flags individual sentences as forward-looking using the same lexicon. Scheme 2 asks the LLM only for a yes/no forward-looking judgment per sentence. Scheme 3 asks the LLM for a structured categorical tag, with discrete labels on horizon, specificity, confidence, and polarity. Scheme 4 asks the LLM to extract factual fields only (modal verbs, time anchors, negation, conditionals, number presence) and then maps those fields to the same categorical attributes as Scheme 3 using deterministic Python rules. Scheme 5 asks the LLM for an ordinal 1–5 score on each axis.

Call-level aggregation. Schemes 3 and 4 produce categorical labels at the sentence level; the corresponding call-level features are shares of forward-looking sentences falling in a target category (e.g., `share_negative` is the fraction of forward-looking sentences tagged `negative`). Scheme 5 produces ordinal scores, and the call-level features are arithmetic means of those scores over the forward-looking sentences in the call (e.g., `mean_polarity`). Both aggregation rules are means, but a share-of-category is the mean of a binary indicator derived after collapsing a multi-class label, while a mean-of-ordinal-score preserves the full 1–5 scale before averaging. This distinction is the precise sense in which Scheme 5 has finer call-level granularity than Schemes 3 and 4.

Model. All production LLM labels use Llama 3.2 3B-Instruct with deterministic decoding (temperature 0). The open-weight model is chosen for reproducibility rather than maximum accuracy. Cross-model validation is run separately with Qwen, Mistral, and Gemma-family models (Section 6.5). Labelling the full corpus required approximately 50 GPU-hours on a single NVIDIA RTX 4090.

5 Empirical Design

The main regression asks whether adverse forward-looking tone predicts different post-earnings drift after positive versus negative earnings surprises. Let $a_{f,t}$ denote a call-level annotation feature oriented so that larger values indicate more adverse forward-looking content, where f index firms and t index earnings-call events (firm-quarters for quarterly reporters, firm-halves for semi-annual reporters). For example, for categorical polarity measures we use the share of forward-looking sentences tagged negative; for averaged ordinal polarity measures, we multiply the mean polarity score by -1 so that larger values again indicate more adverse tone. We standardize this feature within firm:

$$\tilde{a}_{f,t} = \frac{a_{f,t} - \bar{a}_f}{s_f}, \tag{1}$$

where $\bar{a}_f$ and s_f are the mean and standard deviation of $a_{f,t}$ within firm f.

The main specification is

$$\begin{aligned} R_{f,t}^{(i,j)} = \alpha_f &+ \beta_+ \, \tilde{a}_{f,t} \mathbf{1}(\mathrm{SUE}_{f,t} > 0) \\ &+ \beta_- \, \tilde{a}_{f,t} \mathbf{1}(\mathrm{SUE}_{f,t} < 0) + \epsilon_{f,t}, \end{aligned} \tag{2}$$

where $R_{f,t}^{(i,j)}$ is the residualized return from trading day i to trading day j after the earnings call, $\mathrm{SUE}_{f,t}$ is the current-quarter standardized unexpected earnings measure, and α_f is a firm fixed effect. The indicators are formed using the sign of the raw current-quarter SUE. Calls with exactly zero SUE do not contribute to either interaction term. Standard errors are clustered by firm.

The primary coefficient is β_+, the slope on adverse forward-looking tone among positive-SUE calls. The directional prediction is $\beta_+ < 0$: when management beats expectations on the headline number but qualifies the outlook with negative, vague, or low-confidence forward-looking language, the beat is less credible as a signal of

Scheme	Name	Sentence-level output	Call-level features
0	Lexicon density	Forward-looking word density over full call	`fwd_density`
1	Lexicon flag	Binary forward-looking flag per sentence	`share_fl`
2	LLM binary	Yes/no forward-looking classification	`share_fl`
3	LLM categorical	Forward-looking, horizon, specificity, confidence, polarity	`share_negative`, `share_vague`, `share_hedged`
4	LLM extract + rules	Extracted fields mapped to categorical attributes	`share_negative`, `share_vague`, `share_hedged`
5	LLM Likert	1–5 ordinal scores on specificity, confidence, horizon, polarity	`mean_polarity`, `mean_confidence`, `mean_specificity`

Table 1: Annotation schemes. The diagnostic varies output granularity while holding the downstream task fixed.

sustained performance, and the stock should underperform over the subsequent window. The coefficient β_- estimates the analogous slope after negative-SUE calls and is reported as a diagnostic asymmetry check. Each row of Tables 2–6 reports an independent fixed-effects regression that varies a single design element (annotation feature, sample, or outcome equation) while holding all others fixed.

For compactness, let R^* denote the residualized post-call return over the primary event window, trading days 6 through 30 after the earnings call (i.e., $R^* \equiv R_{f,t}^{(6,30)}$). This window is used throughout the paper to avoid choosing different windows for different annotation schemes or samples. Figure 2 visualizes the US validation pattern across the reference window and window sweep. Appendix F reports the full window-sweep design.

We also use a substitute face-validity test that replaces post-call returns with next-quarter fundamentals:

$$\begin{aligned}\mathrm{SUE}_{f,t+1} = \alpha_f + \gamma\,\tilde{\mathrm{SUE}}_{f,t} \\ + \delta_+\,\tilde{a}_{f,t}\mathbf{1}(\mathrm{SUE}_{f,t} > 0) \\ + \delta_-\,\tilde{a}_{f,t}\mathbf{1}(\mathrm{SUE}_{f,t} < 0) + \epsilon_{f,t},\end{aligned} \tag{3}$$

where $\tilde{\mathrm{SUE}}_{f,t}$ is current-quarter SUE standardized within firm. The prediction is $\delta_+ < 0$: after a positive current-quarter surprise, more adverse forward-looking tone should predict a weaker next-quarter earnings surprise. Table 8 summarises all notation used.

6 Results

6.1 US corpus: the annotation spectrum recovers a PEAD signal

Table 2 reports the main US regression. The strongest drift result is Scheme 4 `share_negative`, with $\beta = -0.0060$ and $t = -2.50$. Scheme 5 low confidence is also significant ($t = -2.20$). Simpler schemes are weaker, with Scheme 0 lexicon density close but below the main LLM result.

Scheme	Feature	β_+	t
0	`fwd_density`	-0.0001	-0.04
1	`share_fl`	-0.0005	-0.27
2	`share_fl`	-0.0020	-0.84
3	`share_negative`	-0.0016	-0.59
3	`share_vague`	-0.0046	-2.06
4	`share_negative`	$\mathbf{-0.0060}$	$\mathbf{-2.50}$
5	$-$`mean_confidence`	-0.0048	-2.20

Table 2: US drift test at R^* ($n = 1{,}789$). Reported coefficient is β_+ in Eq. 2. Rows are selected from the full scheme-feature grid. Bold marks the strongest hypothesis-consistent row, discussed in the main text.

Feature	R^*	SUE_{t+1}
Scheme 3 `share_negative`	-0.59	$\mathbf{-3.06}$
Scheme 4 `share_negative`	$\mathbf{-2.50}$	-1.05
Scheme 5 $-$`mean_confidence`	-2.20	-2.68
FinBERT `share_negative`	$+1.01$	-2.84
FinBERT `net_sentiment`	$+0.20$	$+3.20$

Table 3: t-statistics for the US comparison across the drift test (Eq. 2, R^* column) and the next-call SUE test (Eq. 3, SUE_{t+1} column). FinBERT detects next-quarter fundamentals but not the drift signal in this sample. Bold marks the strongest hypothesis-consistent (negative) row in each column.

The next-quarter SUE test supports the interpretation that the labels contain forward-looking information rather than only return noise. Scheme 3 `share_negative` predicts next-quarter SUE with $t = -3.06$, and Scheme 5 low confidence predicts next-quarter SUE with $t = -2.68$. Scheme 4 is weaker on this fundamentals test ($t = -1.05$) but strongest on the return-drift test. This difference is useful rather than problematic, because the schemes are not interchangeable measurements of a single latent factor.

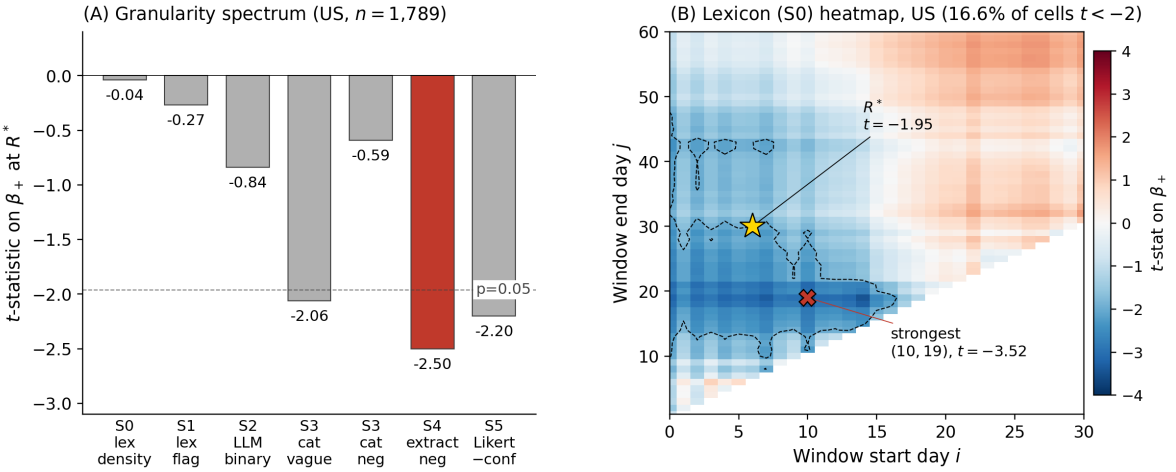

Figure 2: US in-distribution validation of the granularity protocol. (A) Coefficient t-statistics on β_+ at the reference window $R^* = R^{6,30}$ across annotation schemes (Table 2). The Scheme 4 LLM extract-plus-rules result (red) is the strongest hypothesis-consistent row. (B) BRB lexicon (Scheme 0) window heatmap on the same US panel, 1,395 cells. The reference window R^* is marked with a star; the strongest cell is at $(10, 19)$ with $t = -3.52$. Dashed contour at $p = 0.05$. 16.6% of cells reach $t < -2$, against a chance rate of 2.5%.

Sample	Obs.	β_+	t
US full	1,789	−0.0060	−2.50
Asia full	2,560	+0.0063	+2.37
Asia non-Cayman	2,117	+0.0033	+1.22
PRC-via-Cayman	443	**+0.0190**	**+2.86**
SGP only	755	−0.0030	−0.93

Table 4: Scheme 4 `share_negative` drift test at R^*. The Asia failure is concentrated in PRC-via-Cayman firms. Bold marks the diagnostic subgroup.

Sector	Firms	β_+	t
Consumer Staples	10	−0.0177	−3.52
Health Care	11	−0.0079	−2.92
Financials	15	−0.0026	−1.25
Information Technology	10	−0.0016	−0.39
Comm. Services	6	−0.0250	−1.72
IT + Comm. Services	16	**−0.0087**	**−1.56**
Full US sample	75	−0.0060	−2.50

Table 5: Selected US sector breakdown for Scheme 4 `share_negative`. The closest US sector analogue to PRC-via-Cayman firms keeps the expected negative sign. Bold marks the row used as the cross-sample comparison.

6.2 Asia corpus: the same categorical signal reverses sign

Table 4 applies the same Scheme 4 feature and the same return window to the Asian panel. The full Asian coefficient reverses sign ($t = +2.37$). The reversal concentrates in the 23 PRC-via-Cayman firms ($t = +2.86$), while the non-Cayman Asian firms are weak and insignificant.

The Asia next-quarter SUE test is null for the LLM schemes. Scheme 4 `share_negative` gives $t = -0.32$, Scheme 3 `share_negative` gives $t = +0.42$, and Scheme 5 low confidence gives $t = -0.51$. FinBERT, in contrast, detects fundamentals in Asia. FinBERT net sentiment gives $t = +3.85$ and FinBERT negative share gives $t = -3.62$ for next-quarter SUE. This matters for interpretation. The Asia failure is not simply that Asian earnings calls contain no forecastable fundamentals. The LLM rung-ladder polarity features fail to transfer cleanly, especially on the drift task and the PRC-via-Cayman subgroup.

6.3 The sign reversal is not a sector artifact

PRC-via-Cayman firms are technology-heavy. To test whether the wrong sign is merely a sector effect, we rerun the US drift regression by GICS sector and on the combined Information Technology plus Communication Services subsample. Table 5 reports the result.

The closest US analogue does not produce a wrong sign. The coefficient is weaker than the full US result but remains negative, which rules out the simplest sector-composition objection. The PRC-via-Cayman reversal is not what technology firms look like under this model.

6.4 Output-format granularity attenuates the failure

The central diagnostic is Table 6. On the same PRC-via-Cayman firms and the same return window, the wrong-sign result is strongest under Scheme 4, weaker under Scheme 3, and weakest under Scheme 5 averaged ordinal scoring. Fig-

Scheme	Output format	β_+	t
3	Categorical JSON tag	$+0.0101$	$+1.47$
4	Extracted fields + categorical rules	$\mathbf{+0.0190}$	$\mathbf{+2.86}$
5	Averaged ordinal 1–5 polarity	$+0.0119$	$+1.13$

Table 6: PRC-via-Cayman drift test by output format. The same LLM and the same firms produce a weaker wrong sign when polarity is averaged-ordinal rather than categorical. Bold marks the row with the strongest wrong-sign result.

ure 3 summarizes the transfer failure and attenuation across output formats.

This result is the paper's main methodological contribution. The failure is not localized to a firm list alone, nor to the broad idea of sentiment. It is localized to a particular annotation design choice, the forcing of sentence-level polarity into categorical bins. Averaged ordinal scoring does not fully fix the problem, but it attenuates it enough to identify output-format granularity as a causal suspect.

6.5 Cross-model agreement and axis independence

We run two checks to ensure the result is not an artifact of one model or one overloaded construct.

First, the labels are not Llama-specific. Rescoring 2,000 sentences from each corpus with Qwen2.5-7B (Qwen et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), and Gemma-2-9B (Gemma Team: Morgane Riviere et al., 2024) under the same Scheme 3 prompt yields four-rater Fleiss' κ (Fleiss, 1971) of 0.50 on polarity in both samples, with all cross-sample deltas below 0.06 (Appendix I). Agreement is moderate rather than high, but the convergence of three independently developed model families argues against a Llama-only labelling artifact.

Second, the signal axes are nearly independent. The maximum pairwise Pearson correlation among negative tone, vagueness, and low confidence is 0.07, and residualizing each axis on the other two attenuates the headline coefficients by less than 5% without changing signs (Appendix J).

6.6 The simple concession-frequency mechanism is not supported

Qualitative inspection suggests that certain PRC-via-Cayman outliers exhibit sentences in which a negative clause is immediately reframed by a positive one. The Alibaba 2020 call provides an illustrative case, juxtaposing "near-term challenges" with "opportunities created by the forces of change" (Table 7, right column). Constructions of this form may plausibly degrade the performance of a sentence-level categorical polarity tagger.

Aggregate evidence, however, does not corroborate a simple frequency-based mechanism. Among Scheme 4 forward-looking negative sentences, the proportion containing concession markers is 32.71% in the US sample, 32.85% among Asia non-Cayman firms, and 32.47% among PRC-via-Cayman firms (Table 10). A two-proportion test comparing PRC-via-Cayman to US firms yields $z = -0.15$. A more discriminating test, defined as the rate at which positive content following a concession marker exceeds positive content preceding it, likewise returns a null result: 45.45% in the US versus 46.95% in PRC-via-Cayman firms ($z = +0.48$) (Table 11).

This null finding warrants careful interpretation. We do not claim that PRC-via-Cayman firms employ concession structures at higher rates. The supported claim is more circumscribed: categorical polarity output fails on this subgroup, and continuous output attenuates that failure. The precise discourse-level mechanism underlying the attenuation remains an open question.

7 Discussion

7.1 Multi-scheme evaluation as a transfer diagnostic

The contribution is a diagnostic method for reference-free annotation transfer. By evaluating a family of annotation schemes rather than one scheme, the method shows when a construct is stable, when it fails, and which design choice is implicated.

The empirical results illustrate why this matters. In the US, Scheme 4 looks like a valid LLM financial-text annotation. It predicts drift, it is directionally aligned with related schemes, and related axes predict next-quarter fundamentals. If we stopped there, the annotation pipeline would look successful. The Asia transfer test reveals a failure that is invisible in the US-only evaluation, and the granularity comparison then shows that the failure is strongest when the model is forced into categorical sentence-level polarity.

7.2 Implications for NLP annotation practice

For NLP researchers using LLMs as annotators, the practical recommendation is simple. Do not validate only one output format. A categorical label, a structured extraction prompt, and an averaged ordinal score can behave differently even

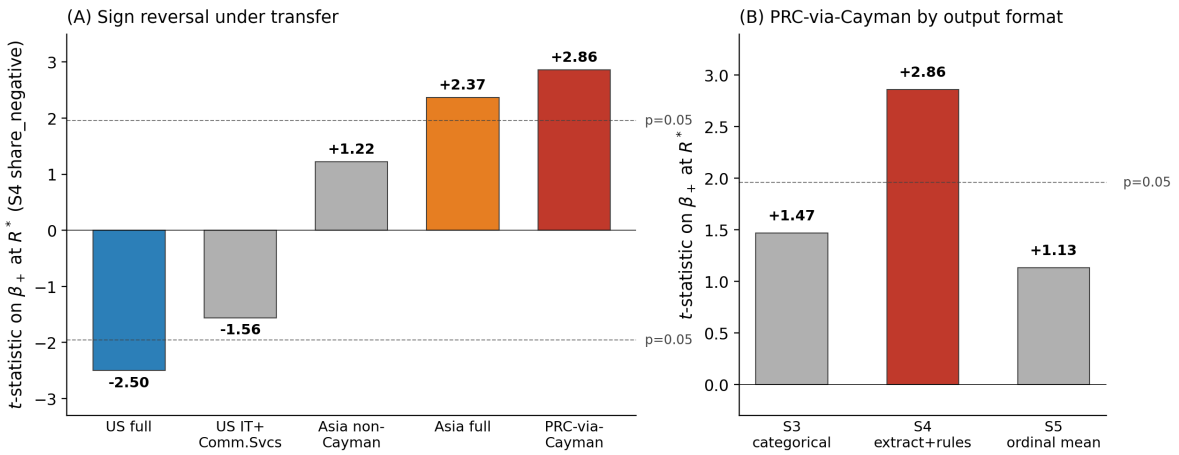

Figure 3: Transfer failure and output-format attenuation. (A) Scheme 4 `share_negative` drift coefficient (t-statistic at R^*) across five samples. The US IT+Comm.Svcs subsample (closest US analogue to PRC-via-Cayman) keeps the expected negative sign, ruling out a sector explanation. The sign reverses on Asia full and concentrates in the 23 PRC-via-Cayman firms ($n = 443$). (B) PRC-via-Cayman drift coefficient across three LLM output formats. Categorical-bin schemes (S3, S4) produce the strongest wrong sign; averaged ordinal scoring (S5) attenuates the failure. Sample sizes match Tables 4–6.

US, correct-sign cell SUE > 0, Scheme 4 fires, $R^* < 0$	**Asia non-Cayman, mixed** SUE > 0, Scheme 4 fires, R^* mixed	**PRC-via-Cayman, wrong-sign cell** SUE > 0, Scheme 4 fires, $R^* > 0$
"On the top line, we're now planning for comparable sales in a wide range centered around a mid-single-digit decline for the remainder of the year." *Walgreens, 2023-08-16;* SUE=+0.96, R^*=−8.3%	"And whilst there will be challenges along the way, China will continue to become more important to the global economy." *Standard Chartered, 2023-02-16;* SUE=+0.62, R^*=−21.0%	"It will present near-term challenges to the development of Alibaba's business across-the-board, but at the same time, we will see opportunities created by the forces of change." *Alibaba, 2020-02-13;* SUE=+1.04, R^*=+28.8%
"While the weakness has moderated somewhat, we anticipate that some softness may continue into the back half of the fiscal year." *Disney, 2023-05-10;* SUE=+0.41, R^*=−9.5%	"And our business has stayed resilient when facing macro challenges, and we are well positioned to benefit from the economy." *Tencent, 2023-03-22;* SUE=+0.55, R^*=−16.0%	"Although we have achieved 39% of our full year paying client growth target in the first quarter, for the time being we'll not adjust our full year target." *Futu Holdings, 2021-05-19;* SUE=+2.90, R^*=+21.9%

Table 7: Illustrative sentences flagged `forward_looking` $\wedge$ `negative` by Scheme 4, drawn from cells in the top decile of $|\tilde{a}_{f,t} \cdot (-R^*)|$ within each subgroup (see Appendix H for construction). Examples are illustrative of how the categorical tagger behaves on subgroup-specific phrasing and are not evidence of differential aggregate concession-marker frequency (Section 6.6, Appendix D).

when they target the same construct. When no gold labels exist, that difference is information. If the schemes agree, the construct is more credible. If they diverge under shift, the divergence is itself a diagnostic.

For financial NLP specifically, the results caution against transferring US-tuned polarity prompts to English-language non-US earnings calls without stress tests. The failure documented here is not solved by choosing a different sector comparison, and it is not revealed by next-quarter fundamentals alone. It appears in the interaction between output format, sample composition, and market-relevant downstream validation.

8 Conclusion

We proposed a reference-free granularity test for LLM annotation transfer. The method evaluates a spectrum of annotation schemes against the same downstream construct and uses the pattern across schemes as the diagnostic. In earnings-call polarity, the test shows in-distribution validity in US firms, a sign-reversed transfer failure in PRC-via-Cayman HKEX firms, and attenuation of that failure when categorical polarity is replaced by averaged ordinal scoring. The supported claim is methodological. Output-format granularity is an important and testable design dimension for LLM-as-annotator pipelines, especially when gold labels are absent and transfer is required.

Limitations

The US sample is modest, with 75 firms and 1,789 main-test observations. The result is statistically meaningful and stable across labels, but a larger S&P 500-scale replication should be done to compare, but not done due to lack of resources.

The PRC-via-Cayman subgroup comprises 23 firms and 443 observations. The fixed-effects estimates are statistically meaningful within this population, but the sample is too narrow to support broad claims about Chinese corporate disclosure more generally. The results should be interpreted as evidence about a specific class of HKEX-listed firms with mainland operating exposure, not as a characterization of Chinese earnings-call language as a whole.

The Asian SUE construction is less precise than its US counterpart because Compustat Global does not expose the announcement-date variables available in Compustat North America. The primary consequence is noise in the fundamentals test rather than in the return-drift test, since the drift test anchors on the call date directly. We note this asymmetry when interpreting the Asia next-quarter SUE results.

The averaged ordinal Likert scheme attenuates the sign reversal on PRC-via-Cayman firms but does not recover the expected negative sign seen in the US sample. The granularity comparison therefore localizes the failure to categorical output format without fully resolving it. Restoring transfer validity would likely require discourse-aware annotation that conditions on intra-sentence contrastive structure, human validation on a stratified cross-market subset, or supervised adaptation of the annotation model to the target disclosure population.

The concession-frequency tests in Appendix D do not reach significance, and the qualitative sentence gallery in Table 7 is accordingly presented as an illustration of how the categorical tagger behaves on subgroup-specific phrasing rather than as evidence for a population-level discourse mechanism. The precise linguistic source of the transfer failure remains an open question.

Acknowledgments

Omitted for anonymous review.

References

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv [cs.CL]*.

Ray Ball and Philip Brown. 1968. An empirical evaluation of accounting income numbers. *J. Acc. Res.*, 6:159.

Victor L Bernard and Jacob K Thomas. 1989. Post-earnings-announcement drift: Delayed price response or risk premium? *J. Acc. Res.*, 27:1.

Wei-Lin Chiang, Joseph Gonzalez, Dacheng Li, Zhuohan Li, Zi Lin, Ying Sheng, Ion Stoica, Zhanghao Wu, Eric Xing, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, and Yonghao Zhuang. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36*, pages 46595–46623. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20:37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186. Association for Computational Linguistics.

J L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76:378–382.

Gemma Team: Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 178 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv [cs.CL]*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. U. S. A.*, 120:e2305016120.

Bas Heerschop, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. 2011. Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv [cs.CL]*.

J R Landis and G G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Joshua Livnat and Richard R Mendenhall. 2006. Comparing the post-earnings announcement drift for surprises calculated from analyst and time series forecasts. *J. Acc. Res.*, 44:177–205.

Tim Loughran and Bill Mcdonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *J. Finance*, 66:35–65.

Vitaly Meursault, Pierre Jinghong Liang, Bryan R Routledge, and Madeline Marco Scanlon. 2022. PEAD.txt: Post-earnings-announcement drift using text. *J. Fin. Quant. Anal.*, pages 1–28.

Jedson Pinto and Federico Siano. 2025. Beyond words: the challenge of measuring financial text across borders. *J. Int. Bus. Stud.*, 56:921–936.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2024. Qwen2.5 technical report. *arXiv [cs.CL]*.

Federico Siano. 2025. The news in earnings announcement disclosures: Capturing word context using LLM methods. *Manage. Sci.*, 71:9831–9855.

Maite Taboada and William C Mann. 2006. Rhetorical structure theory: looking back and moving ahead. *Discourse Stud.*, 8:423–459.

Paul C Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *J. Finance*, 62:1139–1168.

Petter Törnberg. 2023. ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv [cs.CL]*.

Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. *Empir Method Nat Lang Process*, pages 162–171.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can ChatGPT reproduce human-generated labels? a study of social computing tasks. *arXiv [cs.AI]*.

A Notation

Symbol	Definition
f	Firm index
t	Earnings-call event index (firm-quarter)
(f,t)	Firm f's t-th earnings call
$a_{f,t}$	Call-level annotation feature, adverse-oriented
$\tilde{a}_{f,t}$	Within-firm z-score of $a_{f,t}$ (Eq. 1)
$\mathrm{SUE}_{f,t}$	Current-quarter standardized unexpected earnings, raw
$\tilde{\mathrm{SUE}}_{f,t}$	Within-firm z-score of $\mathrm{SUE}_{f,t}$
$\mathrm{SUE}_{f,t+1}$	Next-call SUE, raw
α_f	Firm fixed effect (scalar, time-invariant)
$\epsilon_{f,t}$	Residual
$R_{f,t}^{(i,j)}$	Residualized return, trading day i to j post-call
R^*	Shorthand for $R_{f,t}^{(6,30)}$
β_+, β_-	Drift-regression slopes (Eq. 2)
$\delta_+, \delta_-, \gamma$	Substitute-test slopes (Eq. 3)
$\mathbf{1}(\cdot)$	Indicator function

Table 8: Notation used in the Empirical Design section.

B Model and Code Licenses

The open-weight models used in this work are subject to the following licenses. **Llama 3.2 3B Instruct** (Meta) is released under the Meta Llama Community License Agreement (`https://www.llama.com/llama3/license/`); it was used for all LLM labelling experiments. **Mistral-7B-Instruct-v0.3** (Mistral AI) and **Qwen2.5-7B-Instruct** (Alibaba Cloud) are both released under the Apache License 2.0 (`https://www.apache.org/licenses/LICENSE-2.0`). **Gemma 2 9B IT** (Google) is released under the Gemma Terms of Use (`https://ai.google.dev/gemma/terms`). The Mistral, Qwen, and Gemma models were used solely for cross-model robustness analysis (Appendix I). **FinBERT** (Araci, 2019) (`ProsusAI/finbert`) do not carry an explicit license declaration on their respective model cards; both are derived from BERT (Devlin et al., 2019), which is released under Apache License 2.0. All models were used for inference only; no fine-tuning was performed.

Code and text contained in the supplementary material are under the MIT License.

C Full US sector breakdown

D Concession-marker robustness

E FinBERT subgroup comparison

The main text reports aggregate FinBERT numbers for the US sample (Table 3) and the full Asia sam-

Sector	Firms	β_+	t
Energy	4	-0.0029	-0.17
Industrials	13	-0.0018	-0.23
Consumer Discretionary	6	$+0.0006$	$+0.12$
Consumer Staples	10	-0.0177	-3.52
Health Care	11	-0.0079	-2.92
Financials	15	-0.0026	-1.25
Information Technology	10	-0.0016	-0.39
Comm. Services	6	-0.0250	-1.72
IT + Comm. Services	16	-0.0087	-1.56

Table 9: Full US sector breakdown for Scheme 4 `share_negative` at R^*.

Sample	FL+neg sent.	Concession rate	Within-call mean
US	10,212	32.71%	33.87%
Asia non-Cayman	7,211	32.85%	32.89%
PRC-via-Cayman	884	32.47%	33.10%
Asia full	8,095	32.81%	32.92%

Table 10: Concession-marker rates among Scheme 4 forward-looking negative sentences. PRC-via-Cayman is not higher than US.

Sample	Concession sent.	Post-marker positive > pre-marker positive
US	3,245	45.45%
Asia non-Cayman	2,253	43.23%
PRC-via-Cayman	279	46.95%

Table 11: Sharper concession-direction test. The PRC-via-Cayman rate is not statistically different from the US rate.

		Drift R^* (t)		SUE_{t+1} (t)	
Sample	n_D/n_S	**net**	**neg**	**net**	**neg**
US full	1789/1715	$+0.20$	$+1.01$	$+3.20$	-2.84
US IT+Comm. Svcs.	376/360	$+0.81$	-0.03	$+3.42$	$\mathbf{-5.24}$
Asia full	2560/2394	-1.40	$+1.47$	$+3.85$	-3.62
Asia non-Cayman	2117/1978	-1.12	$+1.35$	$+3.61$	-3.42
PRC-via-Cayman	443/416	-0.77	$+0.62$	$+1.20$	-1.00

Table 12: FinBERT subgroup decomposition. "net" is call-level net sentiment (positive minus negative softmax probability); "neg" is the share of sentences classified as negative. n_D is the drift-test sample size; n_S is the next-call SUE sample size. Reported figures are t-statistics on β_+ in Eq. 2 (drift columns) and δ_+ in Eq. 3 (SUE columns). Bold marks the cell discussed in the second paragraph.

ple (Section 6.2). This appendix extends those numbers to the three subgroups used in the LLM analysis: the US Information Technology plus Communication Services subsample, the Asia non-Cayman subsample, and the PRC-via-Cayman subgroup. The motivation is that the central diagnostic claim in Section 6.4 attributes the wrong-sign result on PRC-via-Cayman drift to LLM categorical polarity output rather than to a general property of finance-domain neural sentiment models.

Testing this claim requires running the same subgroup decomposition through a non-LLM, finance-pretrained baseline. FinBERT (ProsusAI/finbert) serves that role. Table 12 reports the subgroup decomposition. Two caveats apply to the comparison. First, FinBERT is run over every sentence rather than only forward-looking sentences, so the FinBERT features measure overall call-level tone, while Scheme 4 measures forward-looking polarity specifically. Second, FinBERT outputs softmax probabilities rather than discrete categorical bins, so a clean Scheme 3-versus-Scheme 5 granularity contrast is not possible within FinBERT. The comparison is therefore informative about whether the cross-sample breakdown generalizes to other AI sentiment models, not about whether granularity matters within FinBERT.

The drift columns show that FinBERT does not replicate the Scheme 4 sign reversal on PRC-via-Cayman. On the same firms and the same return window, Scheme 4 `share_negative` gives $t = +2.86$ (Table 4), whereas FinBERT `share_negative` gives $t = +0.62$ and FinBERT `net_sentiment` gives $t = -0.77$. The FinBERT drift coefficients are insignificant across every sample, and the strongest wrong-sign cell ($t = +1.47$ on Asia full `share_negative`) is well below conventional thresholds. This is consistent with the Section 6.4 interpretation. A categorical sentence-level polarity tagger that forces every forward-looking sentence into a discrete bin produces a sharp wrong-sign coefficient on PRC-via-Cayman, while a softmax-classifier baseline that does not condition on forward-looking status simply attenuates to noise on the same firms.

The next-call SUE columns tell a complementary story. FinBERT detects next-quarter fundamentals across every sample except PRC-via-Cayman, with $|t| > 2.84$ in the four non-Cayman cells and $|t| < 1.21$ in the PRC-via-Cayman pair. Two implications follow. First, FinBERT does pick up forward-looking information from Asian English-language earnings calls, supporting the Section 6.2 reading that the LLM Asia drift failure is not driven by the absence of forecastable information in the underlying text. Second, the PRC-via-Cayman degradation is shared between FinBERT and the LLM schemes on the SUE task even though it is not shared on the drift task, so the drift result is the cleaner diagnostic for the output-format claim. A side observation worth flagging is the US IT plus Communication Services row, where FinBERT `share_negative` predicts next-call SUE with $t = -5.24$, stronger than the full US sample. The LLM Scheme 4 result on the same sector is much weaker ($t = -0.39$, Appendix C), suggesting that even within the US the categorical LLM

scheme is relatively weakest in the sector where a non-LLM transformer baseline is strongest.

F Window heatmap design

For each pair (i, j) such that $0 \leq i < j \leq 90$, the main regression is estimated using residualized return from trading day i to trading day j after the call. The resulting grid contains 1,395 cells.

G Prompt summaries

Scheme 2 asks whether a sentence is forward-looking and requires a yes/no answer. Scheme 3 asks for a JSON object containing forward-looking status, horizon, specificity, confidence, and polarity. Scheme 4 asks the model to extract factual fields only, including number presence, time anchor, modal verbs, negation, conditionals, and whether the subject is a management claim, after which deterministic rules map the extracted fields to polarity and related features. Scheme 5 asks for 1–5 scores for specificity, confidence, horizon, and polarity.

H Sentence gallery: construction and additional examples

This appendix documents how the illustrative sentences in Table 7 were selected, and provides additional examples that did not fit in the main text. The exercise is descriptive. The aggregate concession-marker tests in Appendix D are the inferential anchor for the claim that PRC-via-Cayman sentences are not differentially concession-heavy at the population level; the gallery is meant to make the failure mode concrete for the reader, not to substitute for that test.

H.1 Bucketing by signal contribution

For each subgroup (US, Asia non-Cayman, PRC-via-Cayman), we restrict attention to positive-SUE calls and compute a per-call contribution score

$$c_{f,t} = \tilde{a}_{f,t} \cdot (-R^*), \tag{4}$$

where $\tilde{a}_{f,t}$ is the within-firm z-score of Scheme 4 `share_negative` and R^* is the within-firm z-score of R^*. A large positive $c_{f,t}$ means a call drove the regression in the hypothesis-consistent direction (high adverse tone, negative subsequent drift); a large negative $c_{f,t}$ means it drove the opposite direction. We bucket calls into `GOOD` (top quintile of $c_{f,t}$), `AVERAGE` (middle 60%), `NOISY` (bottom quintile), and `OUTLIER` (top 2% by $|c_{f,t}|$). Within each bucket we draw up to four calls and, per call, up to three forward-looking negative sentences (Scheme 4) of moderate length (40–350 characters) to keep examples readable.

H.2 Descriptive observation: GOOD-bucket yield by subgroup

Subgroup	Pos-SUE calls	GOOD examples sampleable
US	1,172	12
Asia non-Cayman	1,663	5
PRC-via-Cayman	274	0

Table 13: Number of forward-looking negative sentences sampled from the top-quintile contribution bucket within each subgroup, under the length filter described above. The PRC-via-Cayman zero is the visible consequence of the sign reversal at the call level: no PRC-via-Cayman call in the top contribution quintile carries Scheme 4 negative forward-looking sentences of moderate length. We report this as a description of how the regression failure manifests at the sentence-sampling level, not as additional inferential evidence.

H.3 Additional examples

The examples below extend Table 7 with sentences from each subgroup's `OUTLIER` and `NOISY` buckets. They are drawn directly from the corpus and are intended to convey the qualitative texture of the cells that drive each subgroup's regression coefficient.

US, additional GOOD-bucket sentences. "On the other side of the ledger, while we're happy to see inflation rates begin to moderate, that's likely to cause some near-term pressure on dollar comps in our frequency categories." (Walgreens, 2023-08-16). "This comparison, coupled with inflationary cost pressures, including from a new union agreement, is expected to drive a modest adverse impact to domestic parks and experiences operating margins." (Disney, 2023-05-10). These sentences carry specific, quantifiable, time-bounded forward warnings, consistent with the kind of content that should predict negative drift.

US, NOISY-bucket sentences (categorical polarity fires but drift is positive). "We expect supply chain headwinds to persist into fiscal 2022, and mitigating them remains our top priority." (Cisco, 2021-11-18; SUE=+2.09, R^*=+8.1%). "So look, we don't know exactly how long the strong market conditions that we're seeing today are going to persist." (ExxonMobil, 2022-04-29; SUE=+0.18, R^*=+12.5%). These are typically references to widely-recognized macro conditions

(supply chains, oil prices) where the negative-tone tag is technically correct at the sentence level but the underlying news is already priced.

Asia non-Cayman, mixed direction. "Profitabilities, we have achieved very high levels in terms of the profit, but I believe that in the second half of this year, we might see more difficulties." (China Reinsurance, 2024-08-30; SUE=+1.21, R^*=+26.9%). "However, stepping back, China CRE has not been a driver of our growth and will not be one in the future." (Standard Chartered, 2023-02-16). These are concrete forward warnings, but the realized return is not negative, which is why the non-Cayman regression coefficient remains insignificant.

PRC-via-Cayman, additional wrong-sign sentences. "We will support our merchants to overcome this challenging time and are taking proactive measures to fulfill our mission to make it easy to do business anywhere." (Alibaba, 2020-02-13). "For example, in the U.S. market, the impact from COVID-19 will continue to weaken." (WH Group, 2021-04-27; SUE=+0.90, R^*=+7.0%). The WH Group sentence illustrates a sentence-level parsing failure that is not a concession structure: "COVID-19 will continue to weaken" refers to the impact receding, which is positive news, but the categorical tagger flags it as negative because of the negative lexical content. Sentence-by-sentence categorical polarity tagging cannot recover the intended polarity here without conditioning on what is weakening.

H.4 Interpreting the gallery

The gallery makes visible that Scheme 4 fires on PRC-via-Cayman sentences whose subsequent drift is the opposite of what the U.S. calibration would predict. It does not show that this happens because of higher aggregate concession-marker frequency, which Appendix D rules out at the population level. Two narrower discourse-level patterns appear in the examples, both of which would defeat a sentence-level categorical polarity tagger: (a) intra-sentence challenge-then-reframe constructions, where the negative clause and the reframing clause co-occur in one sentence (e.g., the Alibaba and Tencent sentences in Table 7); and (b) sentence-level parsing failures on phrases that contain negative lexical content but refer to the recession of a negative state (e.g., the WH Group example). These are candidate mechanisms for future work and not claims supported by the present aggregate tests.

I Cross-model agreement: protocol and full results

Section 6.5 reports that the LLM annotation labels are not an idiosyncratic property of any single open-weight model, and that this property holds in both the US and Asian corpora. This appendix documents the protocol behind those claims and presents the full pairwise and multi-rater agreement tables.

I.1 Protocol

We draw a random sample of $N = 2{,}000$ sentences from each corpus (US and Asia) from the Scheme 3 labelled data, with fixed seed for reproducibility. Each sentence is re-scored with three additional open-weight instruction-tuned models, holding the prompt, decoding parameters, and JSON parser identical to the production run: Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, and Gemma-2-9B-IT. All challengers run at temperature 0 in 4-bit NF4 quantisation. The baseline labels come from Llama-3.2-3B-Instruct, the model used for all production annotation in the paper.

Each model returns the five Scheme 3 categorical fields per sentence: `is_forward_looking`, `horizon`, `specificity`, `confidence`, and `polarity`. Following the Scheme 3 specification, the four content axes are valid only when `is_forward_looking` is true; non-forward-looking sentences receive null values on those axes. The sample size used to estimate agreement on each axis therefore depends on the rater pair: it is the number of sentences for which both models in the pair set `is_forward_looking` to true.

I.2 Agreement statistics

We report three quantities per axis. *Pairwise Cohen's* κ (Cohen, 1960) measures agreement between two raters relative to chance, defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{5}$$

where p_o is observed agreement and p_e is expected agreement under independence of the two raters' marginal label frequencies. *Fleiss'* κ (Fleiss, 1971) extends this construction to four raters by replacing the pairwise observed-agreement term with the mean over all $\binom{4}{2}$ pairwise agreements per item,

and the chance term with the squared category-frequency mass across raters. *Four-way exact-agreement rate* is the unadjusted proportion of items on which all four models return the same label; it does not correct for chance and is reported only as an interpretive anchor.

Both kappas drop items with missing values on either rater (Cohen's) or on any rater (Fleiss'). For the four content axes this means the agreement estimate is conditioned on raters agreeing that the sentence is forward-looking, which is the relevant population since downstream features (`share_negative`, `share_vague`, etc.) are only computed over forward-looking sentences. We interpret κ values against the conventional thresholds in Landis and Koch (1977): 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial.

I.3 Cross-model Results

Tables 14 and 15 report the full agreement structure for the US and Asian samples respectively. Three patterns are visible in both samples.

First, agreement on every content axis falls in the moderate range under the four-rater Fleiss' κ: polarity (0.50 US, 0.50 Asia), specificity (0.47, 0.48), and confidence (0.44, 0.50). The forward-looking gate is also moderate (0.45, 0.41). Horizon is the weakest axis in both samples (0.36, 0.38), and we do not rely on horizon-derived features in the main text.

Second, pairwise agreement is not uniform across model pairs. Qwen and Gemma show the highest agreement on three of five axes in each sample (polarity, the forward-looking gate, horizon), while pairs involving Llama and pairs involving Mistral are systematically lower. This is consistent with model-family clustering: Qwen and Gemma are both relatively recent and trained on overlapping data distributions, and their joint behaviour is closer than either is to the Llama or Mistral baselines.

Third, agreement is stable across the US and Asian samples. The maximum absolute Fleiss-κ delta across the five axes is 0.057 (confidence, $0.439 \rightarrow 0.496$). Polarity, the axis that drives all downstream regression results, has a delta of 0.006. This stability is what we mean by the main-text claim that the cross-model check does not produce qualitatively different conclusions on the two corpora.

The agreement is moderate rather than high, which we read as a constraint on what the cross-model check can establish. It does not prove that the LLM labels are correct in any absolute sense, since the four models could share systematic errors inherited from overlapping pre-training corpora. It does argue against the alternative interpretation that the downstream return-drift result, including the sign-reversed Asian result, is driven by a Llama-specific labelling quirk: three independently developed model families recover similar enough labels in both samples that a Llama-only artifact is implausible as an explanation for the cross-corpus transfer failure.

J Axis independence: protocol and full results

A natural concern is that the three LLM-extracted signal axes used in the main text, namely Scheme 3 `share_vague`, Scheme 4 `share_negative`, and Scheme 5 $-$`mean_confidence`, proxy for a single latent "bad-news" factor. If so, any reported separation between drift and next-quarter SUE results would be three measurements of the same underlying construct rather than evidence for distinct linguistic dimensions. This appendix reports the two tests that rule this out.

J.1 Protocol

We work on the US matched panel ($N = 1{,}789$ for the drift test, $N = 1{,}715$ for the next-call SUE test). Each axis is first standardized within firm following Eq. 1. Pairwise Pearson and Spearman correlations are computed at the call level. Confidence intervals are obtained from a block bootstrap that resamples *firms* with replacement (2,000 iterations, fixed seed), preserving within-firm correlation structure. We compute pairwise correlations not only for the three signal axes but also for six related axes from the same rungs (Scheme 3/4 `share_hedged` and `share_positive`, Scheme 5 `mean_specificity` and `mean_polarity`) to map where the matrix is dense and where the signal axes sit relative to that density.

The residualization test then orthogonalizes each signal axis against the other two. For each signal axis $a^{(k)}_{f,t}$, $k \in \{\text{vague}, \text{negative}, -\text{confidence}\}$, we fit

$$\tilde{a}^{(k)}_{f,t} = \alpha_f + \sum_{\ell \neq k} \beta_\ell \tilde{a}^{(\ell)}_{f,t} + r^{(k)}_{f,t},$$

within-firm and use the residual $r^{(k)}_{f,t}$ in place of $\tilde{a}^{(k)}_{f,t}$ in Eq. 2 (drift) and Eq. 3 (next-quarter SUE). Bootstrap CIs on the resulting t-statistics use the same

	Pairwise Cohen's κ						Multi-rater	
Axis	L–Q	L–M	L–G	Q–M	Q–G	M–G	**Fleiss κ**	**4-way agree**
`is_forward_looking`	0.418	0.388	0.314	0.474	0.689	0.402	0.450	0.704
`horizon`	0.235	0.230	0.251	0.565	0.611	0.437	0.362	0.244
`specificity`	0.520	0.393	0.373	0.441	0.512	0.570	0.466	0.458
`confidence`	0.452	0.353	0.423	0.524	0.518	0.596	0.439	0.489
`polarity`	0.468	0.448	0.451	0.472	0.687	0.545	0.501	0.493

Table 14: Cross-model agreement on Scheme 3 categorical labels, **US sample**, $N = 2{,}000$ sentences. L = Llama-3.2-3B-Instruct (baseline), Q = Qwen2.5-7B-Instruct, M = Mistral-7B-Instruct-v0.3, G = Gemma-2-9B-IT. Each pairwise cell drops sentences on which either model labelled the sentence non-forward-looking, so pairwise n varies by axis and pair (forward-looking gate: $n \approx 2{,}000$; content axes: n between 182 and 1,873). Fleiss' κ and the four-way agreement rate condition on all four models marking the sentence as forward-looking (n between 137 and 1,998).

	Pairwise Cohen's κ						Multi-rater	
Axis	L–Q	L–M	L–G	Q–M	Q–G	M–G	**Fleiss κ**	**4-way agree**
`is_forward_looking`	0.277	0.259	0.221	0.489	0.664	0.419	0.406	0.743
`horizon`	0.213	0.221	0.270	0.627	0.558	0.486	0.375	0.268
`specificity`	0.496	0.410	0.384	0.462	0.552	0.531	0.482	0.483
`confidence`	0.456	0.428	0.556	0.532	0.561	0.591	0.496	0.556
`polarity`	0.501	0.418	0.447	0.476	0.664	0.529	0.495	0.479

Table 15: Cross-model agreement on Scheme 3 categorical labels, **Asian sample**, $N = 2{,}000$ sentences. Column abbreviations and conditioning as in Table 14. Pairwise n for content axes ranges from 165 to 1,853; Fleiss' κ n ranges from 124 to 1,999.

firm-block resampling scheme. If the three axes share substantial variance, residualization should attenuate the coefficients; if they capture genuinely distinct linguistic features, residualization should leave the coefficients close to their raw values.

J.2 Pairwise correlations

Table 16 reports the pairwise correlations among the three signal axes. All absolute Pearson values are below 0.08, the maximum implied shared variance is below 0.01, and two of the three Spearman CIs include zero.

Axis pair	**Pearson [95% CI]**	**Spearman [95% CI]**
Vague vs. Negative	$-0.06\,[-0.11, +0.01]$	$-0.05\,[-0.10, +0.01]$
Vague vs. −Confidence	$+0.07\,[+0.01, +0.16]$	$+0.06\,[-0.02, +0.13]$
Negative vs. −Confidence	$+0.05\,[+0.00, +0.11]$	$+0.08\,[+0.02, +0.13]$

Table 16: Pairwise correlations among the three LLM signal axes, within-firm z-scores, with block-bootstrap 95% CIs (2,000 iterations, resampling firms). $N = 1{,}789$ call-level observations.

Figure 4 extends the matrix to nine axes drawn from Schemes 3, 4, and 5. The pattern is consistent with the axes measuring distinct features rather than a single latent tone factor. Within-scheme axes correlate where they should: Scheme 5 polarity and Scheme 5 specificity correlate -0.43 with −confidence, reflecting their shared Likert extraction; Scheme 3 `share_positive` and `share_negative` correlate -0.66, which is near-mechanical for category shares of a near-zero-sum polarity tag. Across schemes, however, the signal axes are essentially uncorrelated, and the strongest off-diagonal cell in the three-axis block remains $|\rho| = 0.07$.

J.3 Residualization

Table 17 reports the raw and residualized t-statistics on the two outcome tests, with block-bootstrap 95% CIs. Figure 5 visualizes the same numbers.

Test, covariate	**t (raw) [95% CI]**	**t (resid.) [95% CI]**
R^* drift, $N = 1{,}789$		
Negative (R4)	$-2.55\,[-3.26, -0.57]$	$-2.43\,[-3.21, -0.46]$
−Confidence (R5)	$-2.24\,[-3.02, -0.28]$	$-1.93\,[-2.75, -0.01]$
Vague (R3)	$-2.11\,[-3.10, -0.20]$	$-2.01\,[-3.10, -0.10]$
Next-quarter SUE, $N = 1{,}715$		
−Confidence (R5)	$-2.74\,[-3.56, -0.57]$	$-2.69\,[-3.53, -0.55]$
Negative (R4)	$-1.08\,[-2.12, +0.67]$	$-0.89\,[-1.97, +0.82]$
Vague (R3)	$-0.48\,[-1.75, +1.18]$	$-0.37\,[-1.67, +1.27]$

Table 17: Raw and residualized t-statistics on β_+ (drift, Eq. 2) and δ_+ (next-call SUE, Eq. 3). Residualized rows use the orthogonalized feature $r^{(k)}_{f,t}$. CIs are block-bootstrap percentiles over 2,000 firm-level resamples.

Two observations follow. First, the axis that dom-

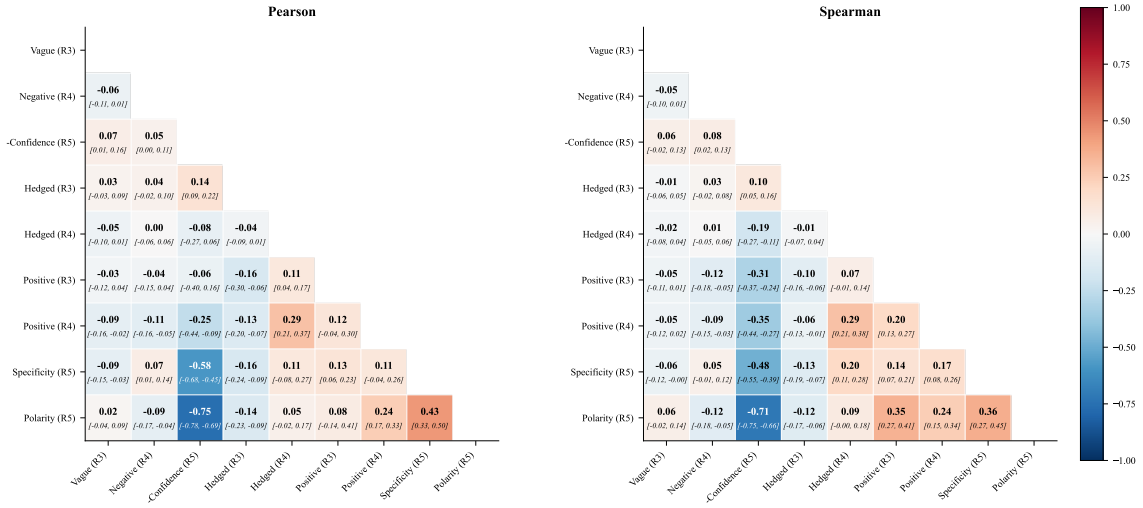

Figure 4: Pearson (left) and Spearman (right) correlation matrices for nine LLM-extracted axes across Schemes 3, 4, and 5. Cell entries show point estimates with block-bootstrap 95% CIs. The three signal axes (`Vague (R3)`, `Negative (R4)`, $-$`Confidence (R5)`) occupy the top-left block and show no pairwise correlation above 0.07.

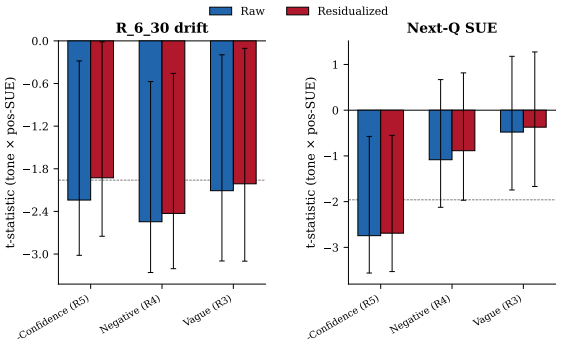

Figure 5: Raw (blue) vs. residualized (red) t-statistics for the three signal axes on the drift test (left) and the next-quarter SUE test (right). Dashed line at -1.96. Whiskers are block-bootstrap 95% CIs. The near-identical bar heights within each pair show that removing all variance shared with the other two axes barely moves the coefficients.

inates each test retains almost all of its t-statistic after residualization: Negative (R4) on drift loses 5% ($-2.55 \rightarrow -2.43$); $-$Confidence (R5) on next-quarter SUE loses 2% ($-2.74 \rightarrow -2.69$). Both residualized CIs exclude zero. The signal carried by each dominant axis is therefore not the variance it shares with the other two axes.

Second, the asymmetry across tests is itself a separability argument. Negative (R4) dominates the return-drift test ($t = -2.43$ residualized) while $-$Confidence (R5) dominates the next-quarter fundamentals test ($t = -2.69$ residualized). A single latent "bad-news" factor expressed three ways would predict both outcomes with comparable strength under any of the three axes. The observed pattern is the opposite: each axis predicts one outcome strongly and the other weakly, and the assignment is consistent with the substantive interpretation that explicit negativity (what is said) maps to short-horizon price reactions while delivery uncertainty (how it is said) maps to future fundamentals.

The bootstrap CIs are wide, reflecting the modest firm count ($N = 75$). The point estimates are nevertheless stable across residualization, and the joint pattern across the two outcome tests rules out a single-factor reading.